Die Hofmethode in der Praxis

Oliver Michel

[redacted]E HOFMETHODE [redacted] DER PRAXIS

Abhandlung
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Universität Zürich

vorgelegt von
Oliver Michel

von Unterseen/BE

Angenommen im Herbstsemester 2012
auf Antrag von
Herrn Prof. Dr. Damian Läge und Frau Prof. Dr. Carolin Strobl

Zürich, 2012

Zusammenfassung

Ausgehend von einem assoziativen Bedeutungsmodell und einem relationalen Textverständnis wird die Bedeutung eines Wortes als Funktion derjenigen Wörter aufgefasst, die das betreffende Wort (wie einen Hof) umgeben. Die Hofmethode greift aus einem längeren Text die fünf Wörter vor und die fünf Wörter nach bestimmten Stichwörtern als «Bedeutungshöfe» heraus und vergleicht diese Höfe mit den Stichwörtern eines anderen Textes. Der paarweise Ähnlichkeitswert zwischen einer Reihe von Wortverwendungen bildet dann die Grundlage einer euklidischen Karte, in der die Ähnlichkeitsrelationen geometrisch dargestellt sind: Texte mit ähnlichen Bedeutungshöfen kommen nahe beieinanderzuliegen. Die aus der Hofmethode resultierende Karte erlaubt eine rasche visuelle und rechnerische Klassifikation dieser Texte, was beispielsweise eine einfache Disambiguierung von Homonymen erlaubt. Suchmaschinen und Übersetzungsdienste profitieren durch die Kontextberücksichtigung: So kann die durch eine Suchmaschine vorgefilterte Resultatsliste in eine relationale Darstellung überführt werden, welche beispielsweise die Orientierung bei eLearning-Produkten erleichtert. Sogar unterschiedlich sprachliche Texte lassen sich in einer einzigen Karte nach ihrer inhaltlichen Bedeutung geordnet darstellen.

abstract

Based on an associative model of semantics and a relational understanding of text we consider the meaning of a word as a function of its surrounding words (the 'halo' or german 'Hof'). The 'Hofmethode' determines the five preceding and the five subsequent words of certain keywords and compares these 'halos' with the keywords of other texts. The resulting matrix of pairwise similarity values can then be used to generate a Euclidean map, in which the similarity relations are represented geometrically: Texts including similar 'halos' will be close together. The resulting map allows a fast visual and computational classification of these texts, which can be used for the disambiguation of homonyms. Search engines and translation services benefit from the context inclusion: the resulting list of a query can be transformed into a relational representation, which facilitates the orientation in a set of e-learning items. Even multilingual texts can be included and semantically ordered in one single map.

INHALTSVERZEICHNIS

Vorwort

Einteilung der Dissertation

Die vorliegende Arbeit gliedert sich in fünf Teile auf. Der erste Teil bietet eine kurze Einführung in das Phänomen der «Bedeutung» aus psychologischer Sicht. Tiefer gehende Gedanken zu Wort und Bedeutung, die hier nur angedeutet werden können, wären spannend und fruchtbar, jedoch gehören sie nicht in den Fokus dieser Arbeit. Durch diese Einführung sollte aber – so hoffe ich – die Neugierde auf das zugrundeliegende Prinzip der Hofmethode geweckt, sowie ein theoretischer, wie auch praktischer Rahmen abgesteckt werden. Ebenfalls im ersten Teil wird die eigentliche Technik der Hofmethode und deren Zusammenwirken mit der NMDS erklärt. Es folgt eine Beschreibung alternativer Techniken und schliesslich eine vorgezogene Bilanz dieser Arbeit: Wo steht die Hofmethode jetzt und wie könnte es weitergehen?

Der zweite Teil widmet sich Techniken, die das Funktionieren der Hofmethode erst ermöglichen, beziehungsweise praktisch greifbar machen: Zu nennen sind hier unter anderem Normierung, Gewichtung und Rauschreduktion.

Auch ohne den Protagonisten des dritten Teils wäre das Arbeiten mit der Hofmethode nicht möglich: Der SemanticMapper, die von uns entwickelte Softwareplattform, welche die Hofmethode implementiert. Der SemanticMapper ist zugleich Demoapplikation, Experimentier- und Testumgebung. Der gesamte dritte Teil enthält eine Aufzählung seiner Funktionen und ist zum Nachschlagen beim Arbeiten mit der Software gedacht.

Der vierte Teil diskutiert ergänzende Verfahren: Auswahl der Keywords, Auswahl bestimmter Texte aus den Karten, Metadaten und – ein Kapitel, auf welches ich mit besonderem Stolz hinweise – Multilinguality.

Der finale fünfte Teil beschreibt verschiedene Experimente und Anwendungen im Zusammenhang mit der Hofmethode. Dieser Teil soll die praktische Nutzbarkeit dieses Verfahrens zeigen und Perspektiven für weitere Forschungsmöglichkeiten eröffnen. Ergänzendes Material findet sich im Anhang.

Illustrationen

Die verschiedenen Teile dieser Arbeit sind durch Trennseiten markiert, die mit Illustrationen und Texten versehen sind. Diese Abbildungen illustrieren in unterschiedlicher Weise die Thematik des

Kontextes und sollen das Thema aus einer weniger technischen Sicht beleuchten. Das mag für eine Doktorarbeit untypisch sein, aber es – nun ja, es passt in den Kontext.

Abkürzungen

Häufige Fachbegriffe werden abgekürzt. Ein Verzeichnis der benutzten Abkürzungen findet sich auf Seite 237.

Dank

Ich möchte vor allem zwei Personen danken – einer aus dem beruflichen, einer aus dem privaten Umfeld. Prof. Dr. Damian Läge unterstützte diese Arbeit durch seinen brillanten Geist, sein enormes Fachwissen und seiner besonderen Fähigkeit motivierend einzuwirken, wenn man in einer Krise steckt, sowie erdend, wenn man leichtfertig über dem Boden der Daten schwebt. Weiter sorgte er für eine technische und finanzielle Infrastruktur, die fruchtbares Forschen erst ermöglichen. Vor allem aber schafft er ein Umfeld, in dem die Freude am Wissen, Entdecken und Erforschen gedeiht. Meine Doktorandenzeit war keine Zeit, die durchzustehen war, um ein Ziel zu erreichen, sondern sie war schon ein Ziel in sich. Forschen ist Freude!

Die zweite Person, der ich danken möchte, ist meine Partnerin Nicole. Sie teilte meinen Enthusiasmus mit der Hofmethode, war Sparring-Partnerin in Diskussionen und gibt mir das Gefühl, ihr spannender Mann zu sein. *Ich reise so gerne mit dir!*

Teil I Das Prinzip der Hofmethode

Plakatkampagne der suva (http://www.suva.ch)

1 Einführung

Der grösste Teil dieser Arbeit beschäftigt sich mit verschiedenen algorithmischen Aspekten der Hofmethode, deren Umsetzung und Anwendung. Dennoch ist es keine rein informationstechnische Arbeit, sondern es ist die Beschreibung einer Anwendung eines universellen Prinzips, dessen Natur psychologischer Art ist. Dieser Umstand ist mir wichtig zu betonen, weshalb ich in diesem einführendem Kapitel – wenigstens holzschnitzartig – darauf eingehen werde.

1.1 Was ist Bedeutung?

Die Hofmethode (HM[1]) ist ein von uns entwickeltes Verfahren, um in automatisierter Weise Wörter nach ihrer inhaltlichen Bedeutung zueinander in Relation zu bringen. Doch was ist überhaupt Bedeutung? Und für wen?

BANK

Was bedeutet dieses Wort? Ist damit ein Geldinstitut gemeint? Oder eine Sitzbank? Wir können es vermuten, aber wir können es nicht wissen.

Die Bank ist frisch gestrichen.

Vor dem geistigen Auge wird man eine frisch gestrichene Sitzbank sehen, wahrscheinlich rot oder grün. Rein semantisch könnte mit «die Bank» aber noch immer ein Geldinstitut gemeint sein. Wieso tendiert man trotzdem eher zur Interpretation «Sitzbank»? Es muss am Kontext liegen. Der Kontext beeinflusst die Bedeutung eines Wortes[2]. Ich möchte sogar so weit gehen und behaupten, dass ein Wort an sich fast keine Bedeutung hat, sondern diese erst durch den Gebrauch in einem spezifischen Kontext erhält. Mit Wittgenstein gesprochen: «Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache» (Wittgenstein, 1960). Schon der magere Kontext in «Die Bank ist frisch gestrichen» reicht, um eine ganz spezifische Vorstellung von Bank zu erhalten. Der Kontext bestimmt die gerade gültige Bedeutung eines Homonyms. Der Kontext erzwingt förmlich ein bestimmtes Wort, weshalb wir falsch geschriebene oder einfach falsch angewendete Wörter trotzdem verstehen. Der Kontext bestimmt die Assoziationen, die uns die Bedeutung eines

1 Ein Abkürzungsverzeichnis befindet sich auf Seite 237.

2 Hörmann (1976) gibt eindrückliche Beispiele, wie der Kontext auch scheinbar unsinnige Wortkombinationen wie «flüssiges Beil» erkärt.

bestimmten Stichworts liefern. Diesen Kontext, also die Wörter vor und nach dem Stichwort, nennen wir den Assoziationshof (daher der Name «Hofmethode»). Wenn wir einen Text lesen, bestimmt also der Assoziationshof die Bedeutung des Stichwortes[3].

Ein Wort, beziehungsweise das Objekt, auf welches mittels dieses Wortes referenziert wird, bekommt so seine jeweilige *assoziative* Bedeutung. Durch häufigen Gebrauch wird der Überlappungsbereich dieser assoziativen Bedeutungen gefestigt und es kristallisiert sich mit der Zeit ein Bedeutungskern heraus, der von den Sprechenden geteilt und akzeptiert wird. Diese assoziative Bedeutung ist nicht zu verwechseln mit der *lexikalischen* Bedeutung[4]: Die Wortmarke «Wald» steht für ein Objekt in der Welt. Das Objekt kann durch bestimmte Eigenschaften definiert werden, welche in einem Lexikon nachzuschlagen sind. Jedoch ist das Wort «Wald» keine Kurzform der Definition. Die Wortmarke «Wald» ist mit Assoziationen aufgeladen, die durch den aktuellen Kontext abgerufen (und gleichzeitig wieder neu gebildet) werden. «Wald» ruft in einem Fachbeitrag zur Forstwirtschaft andere Assoziationen hervor, als der «Wald» des Rotkäppchens.

Bisher wurde im informationstechnischen Umfeld vorwiegend mit der lexikalischen Bedeutung hantiert, jedoch wird man so dem vollen Bedeutungsumfang eines Wortes nicht gerecht. Die Natur der Sprache ist lebendig und dynamisch – die assoziativen Bedeutungen von Wörtern verändern sich über die Zeit; über die Sprecher; über die Situation. Die Berücksichtigung des Assoziationshofes ist unerlässlich, will man mit natürlicher Sprache arbeiten. Das Wort selbst kann ohne Assoziationshof sogar weitgehend nichtssagend sein, weshalb Osgood (Osgood, 1990; Originalartikel in Osgood, 1963) in einem ähnlichen Zusammenhang von der «meaningless form *play*» (S. 114) spricht. «Play» kann als Verb, Nomen oder Adjektiv gebraucht werden, in jeder Wortart wiederum mit verschiedenen Bedeutungen. In einem Gedankenexperiment entwirft Osgood dann ein Verfahren (Semantic Key Sort), um die Bedeutung eines Wortes aus dessen Umfeld heraus zu lesen. Die Wörter im Kontext erhöhen, beziehungsweise reduzieren demnach die Wahrscheinlichkeiten bestimmter Bedeutungen – oder anders ausgedrückt: Erlernte Bedeutungseindrücke werden durch den Kontext abgerufen.

Sogar der Aufbau der lexikalischen Bedeutung kann rein über den Kontext erfolgen. Was für Osgood ein Auswuchs seiner Zeit war (er forschte in den 1950er bis 90er Jahren), ist für uns vermutlich alltäglicher: Neue Begriffe (Objekte) werden nur mittels Wörter erlernt, die andere für uns geschrieben haben. Osgood nennt diese Begriffe «assigns» (S. 330), als Beispiele führt er «Ho Chi Minh», «megatons» oder «Vietcong» an, für uns aktueller sind «Vogelgrippe», «Al Kaida» oder – aus der Lokalpresse – «gekröpfter Nordanflug».

3 Natürlich sind unsere Assoziationen nicht alleine von Text geprägt. Unser Kognitionsapparat arbeitet multimodal; wir spüren die Wärme der Sonne und riechen eine Wiese. Hörmann weist zudem auf den situativen Kontext hin: «Was erwarten wir in einer Situation?»

4 Auf die ursprüngliche Aneignung des lexikalischen Wissens wird hier nicht eingegangen.

Nicht nur die assoziative Bedeutung, sondern auch die komplette Vorstellung über die Natur dieser Objekte wurde nicht durch direkte Erfahrung mit der Realität erstellt, sondern entstammt den Wort-assoziationen anderer.

Wortassoziationen können im Laufe der Zeit ändern. Ein weiteres «assign» von Osgood ist «Cuba». Für ihn dürften damit Sowjets, Raketen und Bedrohung assoziiert gewesen sein, für uns eher Salsa, Mojito und Buena Vista Social Club. Auch das eingangs erwähnte Wort «Bank» hat in den letzten Jahren bedingt durch die Finanzkrise eine assoziative Wandlung erfahren.

Lassen sich diese Wortassoziationen über die Assoziationshöfe maschinell erfassen? Reicht dieser Assoziationshof aus, um eine systematische Bedeutungsvarianz zu erzeugen? Falls ja, dann könnten wir den Hof eines Stichwortes in Text A mit dem Hof desselben Stichwortes in Text B vergleichen. Wären sich die Höfe ähnlich, dann wären sich auch die Bedeutungen ähnlich.

1.2 Bedeutung für wen?

Wortbedeutung ist nicht nur dynamisch, sie ist auch subjektiv – sie braucht ein Subjekt. Ein Subjekt muss eine Situation interpretieren, um eine Bedeutung für sich herleiten zu können.

Maturana und Varela (Maturana & Varela, 1990) charakterisieren das Systembild der «Autopoiesis» (selbsterschaffende Systeme): Lebewesen zeichnen sich demnach dadurch aus, dass sie mit ihrer Umwelt interagieren müssen, um ihre Homöostase zu wahren. Das System ist dabei durch «strukturelle Koppelung» mit der Umwelt verbunden – das Milieu löst Strukturveränderungen in der Einheit aus. Dabei existiert für die Einheit nur diejenige Umwelt, die das System beeinflusst. Das System erschafft sich quasi diese Umwelt durch die Art und Weise, wie es durch sensomotorische Beziehungen damit verwoben ist – die Umwelt existiert nicht «da draussen», sondern wird durch das System generiert. Maturana und Varela beschreiben das Beispiel eines U-Boot-Kapitäns, der in seinem U-Boot geboren wurde, nur die Instrumente kennt und weiss, wie er auf Änderungen in den Anzeigen reagieren muss. Die Kognition eines Organismus ist «ein Tun im Sinne sensoeffektorischer Korrelationen in den Bereichen von Strukturkoppelung, in denen er existiert».

Dieses Prinzip setzt sich auch in der Sprache fort (S. 228):

> Indem [das sprachmächtige Wesen] in der Sprache mit anderen Beobachtern operiert, erzeugt dieses Wesen das Ich und seine Umstände als sprachliche Unterscheidungen im Rahmen seiner Teilnahme an einem sprachlichen Bereich. Auf diese

> Weise entsteht Bedeutung (Sinn) als eine Beziehung von sprachlichen Unterscheidungen. Und Bedeutung/Sinn wird Teil unseres Bereiches der Erhaltung der Anpassung.

Interessanterweise finden sich hier Parallelen zu Osgood, der mit dem «Semantischen Differenzial» in seinem berühmten Artikel «The Nature and Measurement of Meaning» ein Verfahren zur konnotativen Bedeutungsmessung von Begriffen fand (Osgood, 1952). Dabei besteht die gesamte konnotative Wirkung nur aus den drei Dimensionen «Evaluation», «Potenz» und «Aktivation». Wie Osgood in (Osgood, 1990) schreibt, gehen diese Dimensionen auf das Nervensystem zurück (S. 189):

> An underlying notion in our research is that these «experiental continua» will turn out to be reflections (in language) of the sensory differentiations made possible by the human nervous system. In other words, it is assumed that discriminations in meaning, which is itself a state of awareness, cannot be any finer or involve any more variables than are made possible by the sensory nervous system.

Bedeutung in der Sprache braucht also ein Subjekt, das eine Situation evaluiert. In der Künstlichen Intelligenz wird das sogenannte «Embodiment» (Pfeifer, 1999) als Grundlage für intelligentes Verhalten gesehen: Erst die Verkörperung eines Subjektes erzeugt die Reibung mit der Umwelt, aus der sich Situationen ergeben, mit denen in angemessener Weise umgegangen werden muss. Voraussetzung für diesen angemessenen Umgang ist die Interpretation, wie Baumeister in einem Buch mit dem ambitiösen Titel «Meanings of Life» beschreibt (Baumeister, 1991). Damit eine Situation eine bestimmte Bedeutung erhält, muss sie verglichen und vernetzt werden (S. 15):

> A rough definition [des Begriffes «meaning»] would be that meaning is shared mental representations of possible relationships among things, events, and relationsships. Thus, meaning *connects* things.

Anhand der «Adaption-level theory» (S. 16) betont er zudem die Relativität von Bedeutung[5].

Peter Krieg[6] (Krieg, 2005) beschreibt diese Mechanismen unter dem Begriff Relationalität: Eine neue Erfahrung wird zu bestehenden in Beziehung gesetzt (S. 137):

> Bezüge oder *Relationen* setzen aber nicht nur Worte in Beziehung zu verschiedenen Kontexten, sie sind unsere eigentliche «Schnittstelle» zur Welt. Denn was immer unsere Sinne erfassen, setzen wir in Beziehung zu anderem: Es schmeckt wie, es sieht aus wie, es gleicht jenem, es unterscheidet sich von diesem ... Wahrnehmung und Sprache machen nichts anderes, als «Eindrücke» in Relationen zu anderen «Eindrücken» zu setzen. Aber da bereits die «Eindrücke» Relationen verkörpern, bleiben als eigentliche Substanz nur die Relationen selbst übrig. Sie sind die eigentlichen Bausteine, aus denen sich unsere erfahrbare Welt aufbaut.

5 Baumeister führt als Beispiel das Monatseinkommen an: Man freut sich über eine Lohnerhöhung und geniesst das zusätzliche Geld. Bald aber gewöhnt man sich daran und man freut sich auf eine weitere Lohnerhöhung, an die man sich jedoch auch wieder gewöhnen wird.

6 Peter Krieg war Journalist und Dokumentarfilmer. In seinem Buch «Die paranoide Maschine» beschreibt er eine neuartige Technik zum Speichern beliebiger Datenmengen. Der Clou dieser Technik ist, dass nicht die Daten selbst, sondern die Relationen gespeichert werden.

Bedeutung ist relativ und relational. Der direkte Bezug zu einer realen Welt ist nicht nötig, aber es braucht ein Subjekt, das interpretiert und Bedeutung für sich herauslesen kann.

Die Hofmethode berechnet Ähnlichkeiten zwischen Worten und Texten, aber sie ist kein Subjekt. Es braucht den menschlichen Leser, der diese Relationen mit Bedeutung füllt.

1.3 Psychologischer Ansatz

Die Grundidee der Hofmethode entspringt also einem psychologischen Verständnis von Sprache (im Gegensatz zu einem technischen, das vor allem die lexikalische Bedeutung eines Wortes nachschlagen will). Für die HM ist der Assoziationshof der Bedeutungsträger eines Stichwortes. Dieser Assoziationshof besteht aus einigen Wörtern vor und nach einem bestimmten Stichwort. Ähnlich, wie das Hirn kontextabhängig mit Informationen umgeht – sie ergänzt, erfindet, interpretiert – rechnet die HM mit einem definierten Kontextbereich. Dieser Ansatz zum maschinellen Umgang mit Sprache erscheint mir – kognitions-psychologisch gesehen – plausibler, als die Sprache mittels eines linguistischen Regelsatzes analysieren zu wollen. Dass dieser Ansatz aber nicht nur psychologisch plausibler, sondern auch technisch fruchtbar ist, wird mit dieser Arbeit gezeigt werden.

1.4 Anwendungsgebiet

Das Hauptanwendungsgebiet der HM liegt sicherlich im Information Retrieval und Data Mining. Das Information Retrieval (IR) ist ein Fachgebiet der Informatik, welches sich der «Informationsrückgewinnung» widmet[7]. Das IR geht davon aus, dass Information in Datenbanken (z.B. dem World Wide Web) zunächst verloren ist und mit geeigneten Techniken wieder beschafft werden muss. IR-Techniken werden vor allem von Internetsuchmaschinen benutzt, kommen aber auch in der Spam-Erkennung und in Textarchivsystemen zum Einsatz.

Das Data Mining hingegen sucht nach neuen Strukturen in einer bestehenden Datenbank. Im sprachlichen Umfeld, beispielsweise in Internetforen, können dies neuartige Themenbereiche sein.

Verschiedene Hürden können verhindern, dass eine gesuchte Information gefunden wird. Am Beispiel von Internetsuchmaschinen werden einige solcher Fälle beschrieben: Eine suchende Person hat eventuell

7 Für einen historischen Überblick und weiterführende Information siehe http://de.wikipedia.org/wiki/Information_retrieval.

nur eine unklare Vorstellung dessen, was sie sucht. In dem Fall möchte sie vermutlich explorierend vorgehen – sie möchte einen Suchbegriff oder sogar einen Beispieltext in eine Suchmaske eingeben und dann die Ergebnisse durchforsten. Wenn sie etwas gefunden hat, das ihrer diffusen Vorstellung genügend entspricht, möchte sie weitere, inhaltlich ähnliche Resultate erkunden oder mit dem gefundenen Text eine neue Suche auslösen.

Ein weiteres Hindernis ist fehlendes Vokabular: Die suchende Person weiss zwar genau, was sie sucht, jedoch nicht, wie die Bezeichnung dafür lautet. Auch in diesem Fall wäre es hilfreich, wenn sie das Gesuchte umschreiben könnte und die Suchmaschine als Antwort Resultate lieferte, die nach inhaltlicher Ähnlichkeit angeordnet wären, damit sich die Suchperson wiederum explorierend weiter tasten könnte.

Das gilt auch umgekehrt – der Suchbegriff kann zwar exakt und richtig sein, die gewünschte Zielseite benutzt aber ein anderes Vokabular. Auch die Synonym- und Homonymproblematik gehört in diesen Bereich: Gleiche Dinge können unterschiedlich ausgedrückt und beschrieben werden und dieselben Begriffe können Unterschiedliches meinen.

Besonders in Archivsystemen begegnet man der Schwierigkeit, dass sich Texte nicht eindeutig kategorisieren lassen. Schlimmer noch: Kategorien können zum Zeitpunkt der Archivierung noch gar nicht existieren. Verwandt damit ist der Bedeutungswandel: Die Bedeutung von Wörtern kann im Laufe der Zeit ändern.

Als letztes Beispiel ist die maschinelle Übersetzung zu nennen: Dass Wörter nicht eins zu eins mittels Lexika übersetzt werden können, wurde hinreichend – in der Praxis meist unfreiwillig – bewiesen. Auch über das Nachschlagen einer Definition lassen sich keine verlässlichen Ontologien bilden. Im Vorwort des DSM-IV (Sass, 1998, S. XVII) heisst es deshalb in einem resigniertem Tonfall:

> Hier wie übrigens auch bei vielen anderen Übersetzungsproblemen haben wir uns mit der allmählich gefestigten Beobachtung zu trösten, dass die Bedeutung vieler Begriffe stärker durch die normative Kraft des Sprachgebrauches als durch definitorische Bemühungen bestimmt wird.

Das Gemeinte lässt sich nicht ohne Kenntnis des Zusammenhangs übersetzen!

Alle diese Schwierigkeiten haben gemein, dass Bedeutung und Wortform zwei sehr unterschiedliche Dinge sind. Salopp ausgedrückt: Wenn der Computer nur wüsste, was ich meine, könnte er mir das Richtige finden! Die Hofmethode springt hier ein, indem sie mit dem Assoziationshof versucht dieses «Meinen» zu erfassen. Durch den Vergleich der Höfe eines bestimmten Stichwortes (in einem oder

mehreren Texten) lassen sich Bedeutungsschwerpunkte eruieren und durch den Vergleich mehrerer Stichworte in einem Texte lässt sich ein Grundthema des Textes finden. Dadurch sollten sich die genannten Probleme zumindest teilweise lösen lassen. Durch die Kontextberücksichtigung sollten Texte nach ihrer inhaltlichen Ähnlichkeit gruppiert werden können, wodurch ein inhaltlich effizienterer Umgang mit Sprache möglich wäre, als dies bei rein lösungsorientierten Ansätzen der Fall ist, die in Kapitel 3 beschrieben werden. Zudem werden durch die Anordnung auf einer semantischen Karte intuitiv (Un-)Ähnlichkeits-relationen sichtbar, die durch eine listenartige Darstellung nicht zu vermitteln sind.

2 Die Hofmethode

Die Hofmethode ist ein mehrstufiges Verfahren, um erst die Höfe (Assoziationshöfe) von Stichworten zu berechnen und anschliessend deren Ähnlichkeiten zu schätzen. Diese Ähnlichkeiten werden in einem weiteren Schritt mittels einer robusten Implementierung der NMDS in einer zweidimensionalen Karte dargestellt. In diesem Kapitel werden zuerst die notwendigen Vorarbeiten und die Bestimmung einiger Parameter beschrieben, anschliessend folgt in Kapitel 2.2 ein konkretes Beispiel, in Kapitel 2.3 das Zusammenspiel mit der NMDS und schliesslich eine Zusammenfassung in Kapitel 2.4.

2.1 Vorarbeiten und Parameterbestimmung

2.1.1 Rauschreduktion

Die geschriebene Sprache enthält Wörter, die wenig Informationswert besitzen. So wie ein menschlicher Zuhörer Wörter nicht versteht oder überhört, können wir aus einem Text redundante Wörter (Stoppwörter) entfernen. Der Einfachheit halber definieren wir die 100 gebräuchlichsten deutschen Wörter[8] als Stoppwörter. Weiter ist (leider) eine Bearbeitung des Textes notwendig, die kein psychologisches Pendant hat, sondern rein rechnerischen Gründen dient: Der Text muss standardisiert werden, d. h. Satz- und Trennzeichen werden entfernt, der gesamte Text wird in Kleinschreibung gesetzt und Ähnliches. Durch diese beiden Massnahmen wird – informationstechnisch gesprochen – Rauschen entfernt.

2.1.2 Bestimmung der Keywords

Je nach Erfordernis wird in der Textbasis nur nach einem bestimmten Stichwort (=Keyword) in sämtlichen Texten gesucht (z. B. wenn man wissen möchte, welche unterschiedliche Ausprägungen ein bestimmtes Wort hat), oder es werden mehrere Keywords in mehreren Texten behoft, um die Ähnlichkeitsbeziehungen ganzer Texte miteinander zu vergleichen. Die Kompilierung der Keywordliste wird in den Kapiteln 14, 15 und 16 besprochen.

8 Eine Zusammenstellung ist auf http://wortschatz.informatik.uni-leipzig.de/ zu finden.

2.1.3 Hofgrösse

Wie gross sollte der Hof sein? Getreu dem psychologischen Ansatz bietet sich Millers «Magische Zahl 7 plus minus 2» (Miller, 1956) an. In verschiedenen Vorversuchen hat sich gezeigt, dass eine Hofgrösse von fünf ausreicht; der Hof soll die fünf Wörter vor und die fünf Wörter nach dem Stichwort umfassen. Diese Grösse ist intuitiv plausibel; wenn man einen Text vor sich hat und ein bestimmtes Stichwort auswählt, erkennt man, dass sehr oft ganz typische (den Sinn des Stichwortes beeinflussende) Wörter in unmittelbarer Nähe zu finden sind. Standardmässig arbeiten wir also mit einer Hofgrösse von fünf, trotzdem könnte eine Anpassung nötig sein, beispielsweise wenn andere Sprachen oder spezifische Textformen untersucht werden.

2.1.4 Gewichtung

Ebenso intuitiv plausibel ist es, dass weiter entfernte Hofwörter an Einfluss auf die Wortbedeutung verlieren, deshalb sollen die Hofwörter in Abhängigkeit zur Distanz zum Stichwort gewichtet werden. Wir entschieden uns für eine Cosinus-Funktion:

$$V = \cos(d \cdot \frac{90}{Hof + 1})$$

«V» ist der Gewichtungswert des Assoziationswortes, «d» die Distanz zum Stichwort, «Hof «ist die Hofgrösse. Die Gewichtung nimmt mit der Entfernung zum Stichwort überproportional ab. Als Resultat erhalten wir für das Stichwort einen 10er-Vektor. Dieser Vektor kann mit dem Vektor desselben Stichwortes eines anderen Textes verglichen werden.

2.1.5 Unscharfe Ähnlichkeit

Die Einträge des Hofes (der Vektor) werden mit denen anderer Höfe verglichen. Damit flektierte Wortformen nicht als ungleich registriert werden, könnte man ein *Stemming* anwenden (s. auch Kap. 9). Als Stemming werden Algorithmen bezeichnet, die ein Wort auf seinen Stamm reduzieren. Dieser Ansatz wäre aber «ingenieursmässig»: Der Text wird so angepasst, dass er maschinell besser verarbeitet werden kann. Hier soll wieder der psychologische Ansatz zum Tragen kommen. Wie erkennen wir zwei ähnliche

Wörter? – Sie sehen ähnlich aus! Also werden bei der Ähnlichkeitsberechnung nicht zwei *gleiche* Wortmarken als gleich behandelt, sondern schon *ähnliche*.

Als Ähnlichkeitsalgorithmus verwenden wir die Levenshtein-Distanz[9]. Diese misst die Schritte, die notwendig sind, um eine Zeichenkette in eine andere zu überführen, wobei die Buchstabenoperationen «Löschen», «Ersetzen» und «Einfügen» erlaubt sind. Diese Distanz korrespondiert nach unserem Ermessen recht gut mit dem optischen Ähnlichkeitseindruck beim Lesen zweier Wörter.

2.2 Ein Hofbeispiel

Unser Textausschnitt A sei folgender (das Stichwort ist bereits fett markiert):

> Bei der Stiefmutter daheim hätte Gretchen nun böse Zeit. Sie gab ihm wieder schwere Arbeiten auf, zankte immerfort und ließ es auch an Schlägen nicht fehlen; ihre dreiäugige Tochter aber arbeitete nichts, sondern putzte sich immerfort und ging ihrem Vergnügen nach. Dennoch war diese nie so schön als Gretchen. Das ärgerte die Stiefmutter, und sie beschloß, es zu verschaffen. Sie führte es tief in einen dichten **Wald** und schickte es dann zu einer Quelle um Wasser. Inzwischen verwandelte sie sich in einen schwarzen Käfer und setzte sich unter einen Strauch; von da wollte sie sehen, wie Gretchen sie suchen und sich verirren solle.

Dieser wird nun der Rauschreduktion unterzogen:

> stiefmutter daheim hätte gretchen nun böse zeit gab schwere arbeiten zankte immerfort liess schlägen fehlen dreiäugige tochter arbeitete nichts sondern putzte immerfort ging ihrem vergnügen dennoch nie schön gretchen ärgerte stiefmutter beschloss verschaffen führte tief dichten **wald** schickte quelle wasser inzwischen verwandelte schwarzen käfer setzte strauch da wollte sehen gretchen suchen verirren solle

Jetzt kann das Stichwort behoft werden. Die Werte werden in eine Tabelle geschrieben.

Tabelle 1: Der Hof des Wortes «Wald» in Text A

Assoziationswort	Wert
beschloss	0.26
verschaffen	0.5
führte	0.71
tief	0.87
dichten	0.97
wald	1
schickte	0.97

9 http://de.wikipedia.org/wiki/Levenshtein-Distanz

Assoziationswort	Wert
quelle	0.87
wasser	0.71
inzwischen	0.5
verwandelte	0.26

Diese Tabelle repräsentiert den Hof des Wortes «Wald» in Text A. Dieser Hof wird nun mit dem Hof desselben Wortes in Text B (hier nicht abgedruckt) verglichen.

Tabelle 2: Der Hof des Wortes «Wald» in Text B

Assoziation	Wert
weges	0.26
dahinzog	0.5
ihn	0.71
just	0.87
dunklen	0.97
wald	1
dichtem	0.97
gebüsch	0.87
vorüberführte	0.71
stiess	0.5
seinem	0.26

Jedes Assoziationswort von Hof A wird mit jedem Assoziationswort von Hof B verglichen. Sind sich die Wörter ähnlich (z.B. «dichten» und «dichtem»: Ähnlichkeitswert = 0.86), werden ihre Werte und der Ähnlichkeitswert multipliziert: $0.97 \cdot 0.97 \cdot 0.86 = 0.81$. Diese Werte werden aufsummiert und bilden den Ähnlichkeitswert der beiden Höfe. Für obige zwei Höfe wäre der aufsummierte Ähnlichkeitswert 0.3. Für jeden Text können mehrere, unterschiedliche Zielwörter behoft werden. Hat man mehrere Texte, werden alle Höfe miteinander verglichen – sofern sich die Assoziationswörter genügend ähnlich sind – und textweise aufsummiert. Diese Ähnlichkeitswerte bilden eine Dreiecksmatrix. Jeder Wert schätzt die Ähnlichkeit zwischen dem jeweiligen Textpaar. Durch eine robuste Form der NMDS lässt sich diese Matrix visualisieren.

2.3 NMDS

Nonmetrische Multidimensionale Skalierung (siehe Borg, 1997 und Mathar, 1997): Dieses Verfahren wird in der kognitionspsychologischen Forschung angewendet, um Daten, in denen es ein «näher zusammen» und ein «weiter weg» gibt, räumlich abzubilden. In der Metapher der Landkarte liegt denn auch der Ursprung des Konzepts der sogenannten *Kognitiven Karten*, die strukturierte Vorstellungen von Ausschnitten unserer Welt widerspiegeln (Marx & Hejj, 1989). Die zugrunde liegende Idee ist, dass die (wahrgenommene oder geschätzte) Ähnlichkeit zwischen Objekten durch die dazwischenliegende Distanz abgebildet wird. Dicht beieinander liegende Objekte sind sich ähnlich, weit voneinander entfernte unähnlich. Ähnlichkeitswerte werden in der Regel durch Multidimensionale Skalierung in der Art in eine niedrigdimensionale, leicht zu interpretierende geometrische Struktur (meistens eine zweidimensionale Karte) überführt, dass die Distanzen zwischen den Objekten deren Ähnlichkeitsrelationen möglichst gut abbilden. Bei der NMDS handelt es sich um ein iteratives heuristisches Verfahren, das daran orientiert ist, die Summe der quadrierten Abweichungen zwischen den Proximitäten und den Distanzen schrittweise zu minimieren; es soll das Stresswert-Minimum für eine Lösung in einem niedrigdimensionalen Raum finden[10].

Wegen ihrer enormen Robustheit wird die NMDS zum integralen Bestandteil der Hofverrechnung: Die Hofmethode schätzt durch den Hofvergleich die Ähnlichkeit zwischen Texten. Diese Ähnlichkeitswerte sind jedoch keine *Messwerte*, die eine postulierte Ähnlichkeit widerspiegeln, sondern es sind Schätzwerte, die die Hofmethode aufgrund mehrerer, psychologisch motivierten Transformationen errechnet. Diese Transformationen führen im Allgemeinen – das zeigt die Empirie der Hofmethode – zu einem sinnvollen Resultat, jedoch kann der konkrete Einzelfall einer Schätzung schlicht und einfach falsch sein. Durch eine unglückliche Formulierung, ein aussergewöhnliches Vokabular, einen knappen Schreibstil oder umständliche Ausschweifungen können Ähnlichkeiten entstehen, wo keine sind und können tatsächliche Ähnlichkeiten übersehen werden. Die NMDS aber beachtet für eine Positionierung auf der Karte nicht nur diesen einen Ähnlichkeitswert, sondern bezieht alle anderen Werte mit ein[11]. Über die Kovarianzen zu den anderen Items relativiert sich eine falsche Schätzung wieder. Damit das Gesamtbild der Karte möglichst gut die Ähnlichkeitsrelationen repräsentiert, ist der Einfluss einzelner Schätzwerte somit gering. Erst dadurch ist es möglich, dass sich beispielsweise zwei Texte in der Karte nahe kommen, die zwar untereinander keine «geschätzten» Ähnlichkeiten aufweisen, jedoch zu denselben Texten Kovarianzen teilen. Die unabhängigen Werte der Dreiecksmatrix werden also untereinander in Relation gebracht.

10 Für eine gut verständliche Einführung siehe Borg, Groenen & Mair (2010).

11 Unser verwendete NMDS-Algorithums (Robuscal) minimiert (sogar eliminiert) darüber hinaus den Einfluss von offensichtlich falschen Schätzungen.

Die NMDS erfüllt somit zwei Aufgaben: Es wird aus grundsätzlich hoch verrauschtem Sprachmaterial ein Optimum an Signal heraus gelesen und es wird uns mit der Metapher der «Karte» eine Visualisierung geboten, dank der wir in intuitiver Weise Strukturen (Cluster) entdecken und so auf einer enorm abstrahierten, verdichteten Informationsebene arbeiten können.

2.4 Die Technik der Hofmethode im Überblick

- Bestimmung der zu untersuchenden Textbasis
- Rauschreduktion (Entfernung von Stoppwörtern, Standardisierung)
- Bestimmung der Keywords
- Hofgrösse bestimmen
- Hofwörter (Assoziationshof) bestimmen und nach der Entfernung zum Keyword gewichten
- Ähnlichkeit der Höfe bestimmen, dabei unscharfen Ähnlichkeitsalgorithmus anwenden
- NMDS rechnen und darstellen (Datenübergabe an ProDaX[12])

12 Der Proximity Data Explorer (Prodax) ist ein statistisches Softwarepacket, das von der Abteilung für Angewandte Kognitionspsychologie an der Universität Zürich entwickelt wurde und einen besonders robusten Algorithmus für die NMDS verwendet. Prodax umfasst zudem weitere strukturentdeckende und visualisierende Verfahren wie Prokrustes-Transformation und Clusteranalyse, siehe auch http://www.prodax.ch.

3 Alternative Techniken

Es existieren verschiedene alternative Techniken im Information Retrieval und Data Mining. Wie auch die Hofmethode bestehen die hier beschriebenen Verfahren aus zwei Schritten: Erst werden Ähnlichkeiten zwischen Texten berechnet, dann werden mittels Kategorisierungsalgorithmen Strukturen erfasst. Die verwendeten Kategorisierungsalgorithmen sind meist das K-Means-Verfahren (Bortz, 1999), Entscheidungsbäume (Quinlan, 1986), Naiver Bayes-Klassifikator (McCallum, 1998) und Support Vector Machines (Joachims, 2001). Wir werden nicht näher darauf eingehen. Die Algorithmen zur Textähnlichkeitsberechnung werden folgend ganz kurz geschrieben.

3.1 Algorithmen zur Textähnlichkeitsberechnung

3.1.1 Vector Space Model (VSM)

Im VSM repräsentiert jedes Wort eines Dokumentes eine Dimension in einem Vektorenraum (Panyr, 1986). Textähnlichkeiten werden aus dem Vergleich zweier Vektorräume berechnet. Die Vektorenräume sind hochdimensional, weshalb die Rechenzeit rasch unpraktisch wird und Dimensionsreduktionsverfahren angewendet werden müssen.

3.1.2 Latent Semantic Indexing (LSI)

Ein solches Verfahren zur Dimensionsreduktion ist das LSI (Deerwester, 1990). Durch Singulärwertzerlegung werden die Dimensionen eines Dokumentes stark reduziert (Deerwester spricht von ca. 100 Dimensionen). Das LSI ist zur Zeit recht populär.

3.1.3 Linear Dirichlet Approximation (LDA)

Die LDA (Blei, 2003) behandelt jedes Item eines Textkorpus als eine Mischung von zu Grunde liegenden Themen. Die Wörter ihrerseits können wiederum verschiedenen Themen zugeordnet werden. Die Ähnlichkeit zwischen Texten besteht aus den Ähnlichkeiten dieser Themen.

Den drei beschriebenen Verfahren ist gemein, dass sie Dokumente als «bag of words» behandeln. Die Reihenfolge der Wörter ist innerhalb der Dokumente völlig irrelevant. Das folgende Beispiel beachtet hingegen die Reihenfolge der Wörter und kommt der Hofmethode somit am Nächsten:

3.1.4 Hyperspace Analogue to Language (HAL)

HAL (Lund, 1996) beachtet – ähnlich wie die Hofmethode – den Kontext eines bestimmten Stichwortes. Ein 10 Wörter langes Lesefenster gleitet über den Text und erfasst das Auftreten von Wortpaaren, wobei auch hier die Paare gemäss ihrer Distanz innerhalb dieses Fenster gewichtet werden. Die Wortpaare werden in eine NxN-Matrix geschrieben. Deren Wert wird erhöht, wenn dasselbe Paar in einem anderen Wortfenster wieder auftaucht. Die Ähnlichkeit zwischen Wörter wird aus dem Winkel im n-dimensionalen Vektorenraum berechnet.

3.2 Eine konkrete Anwendung: Das Jam-Framework

Exemplarisch wird eine State-of-the-Art-Anwendung – das Jam-Framework – von IBM vorgestellt, das einen etablierten Algorithmus implementiert.

Das Jam-Framework von IBM darf ruhig als State-of-the-Art-Anwendung beschrieben werden, demonstrierte IBM doch im Februar 2011 eindrücklich die Leistungsfähigkeit ihrer «Wissensmaschine» Watson[13]. Das Jam-Framework wurde von IBM entwickelt, um in online abgehaltenen Diskussionen (sogenannten Jams) neue Themen (Threads) in Echtzeit zu detektieren (Spangler et al., 2006). Dabei werden keine neuartigen Algorithmen verwendet, sondern bestehende Techniken des Data Minings zusammengefasst (mit dem COBRA[14]-Toolset) und einem menschlichen Analysten unterstützend zur Verfügung gestellt, damit dieser allfällige Erkenntnisse noch während des Jams zurückmelden kann (beispielsweise auf eine besonders innovative Idee aufmerksam machen, die in einem Thread lebhaft diskutiert wird). Spangler nennt das «interactive text mining» (S. 787).

Mittels Wortfrequenzen wird ein Vector Space Model gerechnet, welches mit dem K-Means-Algorithmus geclustert wird. Damit der Analyst die Bedeutung der resultierenden Cluster erfassen kann, werden diese nach dem – in diesem Cluster – dominierenden Begriff benannt. Weiter wird jeder Cluster

13 http://www-03.ibm.com/innovation/us/watson/index.html
14 http://www.almaden.ibm.com/asr/projects/cobra/

beschrieben, indem die Wortfrequenzen relativ zum Cluster und relativ zum Gesamtdatenbestand angezeigt werden, sowie eines ID3-Entscheidungsbaumes (Quinlan, 1986).

Auch eine visuelle Darstellung steht dem Analysten zur Verfügung. Da der Vektorraum der Wortfrequenzen zu hoch dimensioniert ist für eine grafische Darstellung, wendet Spangler die CViz-Methode (Dhillon, 1998) an. Dabei wird mit den Zentroiden dreier Kategorie-Cluster anhand ihrer Wortfrequenzen eine Ebene aufgespannt, in der die einzelnen Dokumente gemäss ihrer eigenen Wortfrequenzen platziert werden können.

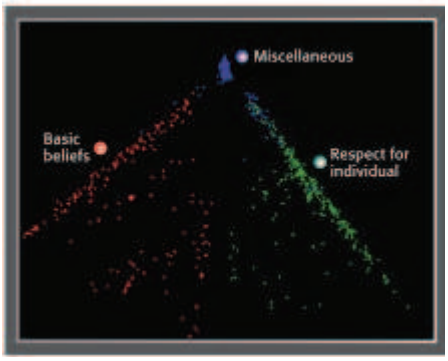


Abbildung 1: Visualisierung von Klassen mit der CViz-Methode

Der Analyst kann die automatisch erstellte Taxonomie nach verschiedenen Kriterien sortieren und verändern. Um neue Themen zu identifizieren, werden die jeweils neuesten Beiträge auf ungewöhnlich hohe Frequenzen von Thermen hin untersucht und allenfalls dem Analysten präsentiert, damit dieser entscheiden kann, ob es sich um eine neue Kategorie handelt.

Die gewonnen Erkenntnisse könne dem Jam in Form einer Themenwolke zurückgemeldet werden:

Abbildung 2: eine vom Jam-Framework erstellte Themenwolke

4 Stand der Dinge

Mit der Hofmethode wurde ein innovatives Verfahren gefunden, konnotative Bedeutungen von Wörtern und semantische Relationen zwischen Texten zu schätzen. Die Hofmethode unterscheidet sich von den im Information Retrieval etablierten Verfahren nicht nur durch die Art und Weise, wie sie diese Ähnlichkeiten berechnet, sondern auch dadurch, dass sie diese Ähnlichkeitswerte als Schätzer für eine Multidimensionale Skalierung nimmt.

Wie die folgenden Kapitel zeigen werden, liegt die Stärke dieser Methode darin, feine semantische Unterschiede in der Verwendung von Wörtern in einer intuitiv verständlichen Weise darzustellen. Der psychologische Ansatz, Wörter im Zusammenhang ihres Kontextes zu betrachten, erweist sich als fruchtbar und wissensgenerierend.

Anwendungsmöglichkeiten gibt es überall, wo digitale Texte nach Inhalt geordnet werden müssen. Hervorzuheben sind die Kategorisierung und Homonymdisambiguierung (Kap. 25, Homonyme: Explorationsexperiment Bach/Golf und Kap. 26, Homonymdisambiguierung: Anwendungsbeispiel); in einem übergeordneten Rahmen die Internetsuche oder maschinelle Übersetzung; auch wissenschaftliche Abstracts (Kap. 22, Bedeutungsähnlichkeiten von Abstracts: Vier Verarbeitungsebenen) und Beschreibungen von e-Learning-Ressourcen (Kap. 23, Projekt edulap) lassen sich hervorragend ordnen.

Im Zusammenhang mit den e-Learning-Ressourcen entwickelten wir ein originelles Verfahren, unterschiedlich-sprachliche Texte ihrer Semantik entsprechend in derselben Karte darzustellen (Kap. 21, Multilinguality). Diese Mehrsprachigkeit ist eine mächtige Lösung und wird in der Praxis sicherlich grossen Nutzen bringen.

Natürlich gibt es Limitationen. Auf die rechnerischen wurde in dieser Arbeit nicht detailliert eingegangen, jedoch ist es unbestreitbar, dass die Hofmethode und die NMDS rechenintensive Verfahren sind, für die man bei grossen Textmengen Speziallösungen einsetzen muss. Auch die Semantik betreffend gibt es Grenzen, die wir zwar ausdehnen, aber nicht aufheben können. So führte beispielsweise das Experiment mit der klinischen Diagnostik (s. Kap. 27, Experiment DSM-IV und Diagnostik) zu keinem in der Praxis verwertbaren Resultat.

Die Hofmethode ist keinesfalls ausgeforscht. Ein spannender Bereich, der künftig stark an Bedeutung gewinnen wird, ist die Verfasseridentifikation. Hier ist nicht die Plagiatsforschung gemeint (auch eine

mögliche Forschungsrichtung), sondern die Autorenschaft von Online-Forumsbeiträgen. In der Terrorismusbekämpfung ist ein Ansatzpunkt, potenzielle Gefahrenherde durch Auswertung von online verfügbarem Material (v.a. Online-Foren) zu identifizieren (für eine aktuelle Übersicht siehe Wiil, 2011). Eine besondere Schwierigkeit ist die Erkennung (als mögliche Attentäter) von Personen, die zwar angesichts ihrer geschriebenen Äusserungen in ein typisches Attentäterprofil passen würden, jedoch keinen Kontakt zu sogenannten «Watchlist Members» haben (Danowski, 2011). Gerade die Hofmethode böte hier ein probates Mittel, Beiträge nach ihrer Semantik einzuordnen und mit bereits klassifizierten Autoren zu vergleichen.

Mit grosser Spannung erwarten wir den praktischen Einsatz der Hofmethode im Projekt edulap und hoffen, dass dieser Algorithmus auch bald in anderen Projekten implementiert werden wird.

Teil II Entwicklung der Hofmethode

Plakatkampagne des Schleudertraumaverbandes (http://schleudertraumaverband.ch)

5 Normierung der Textähnlichkeitswerte: SharedTargetWords vs. TotalTargetWords

5.1 Überblick

SharedTargetWords und TotalTargetWords sind unterschiedliche Arten Textähnlichkeitswerte zu normieren. Dieses Kapitel vergleicht deren Auswirkungen mit derjenigen einer simplen Aufsummierung. Zwar sind die Ergebnisse nicht homogen, jedoch spricht die theoretische Herleitung für TotalTargetWords.

5.2 Einleitung

Werden die Keywords von Texten mittels Hofmethode verglichen, entstehen Ähnlichkeitswerte. Werden diese Werte aufsummiert, kann der entstehende Wertebereich sehr weit sein und die relationalen Beziehungen der Texte untereinander verzerren. Deshalb setzt man eine Normierung ein, die die Vergleichswerte in einen engeren Bereich bringt, bevor sie in die Dreiecksmatrix (DEM) kommen.

5.2.1 SharedTargetWords

SharedTargetWords basiert auf der Schnittmenge der gefundenen TargetWords in den beiden zu vergleichenden Texten: Alle Hofvergleiche werden aufsummiert und durch die Anzahl *gemeinsamer* Keywords geteilt. In anderen Worten: Ausschlaggebend für den Textähnlichkeitswert ist die durchschnittliche Ähnlichkeit der gemeinsamen Höfe.

5.2.2 TotalTargetWords

TotalTargetWords basiert auf der Gesamtanzahl aller gefundenen Keywords der beiden Texte. Der Ähnlichkeitswert eines gemeinsamen Zielwortpaares wird durch die Anzahl gefundener Zielwörter geteilt. Somit sind sich zwei Texte ähnlicher, wenn mehrere gemeinsame Zielwörter vorkommen, als wenn nur ein einziges Paar gefunden wird, dessen Wert aber relativ hoch ist – das ist inhaltlich plausibler.

Folgende Tabelle zeigt anhand eines fiktiven Beispiels, wie sich die verschiedenen Normierungen auf die Textähnlichkeitswerte auswirken:

Tabelle 3:Auswirkungen der verschiedenen Normierungen

	Text A	**Text B**	**Text C**
gefundene Zielwörter pro Text (n)	20	30	40
Anzahl gemeinsamer Zielwörter (n_{shared})	4 (1.2 + 1.2 + 1.2 + 1.2)	1 (1.3)	
Summe (S) der gefundenen Ähnlichkeitswerte (=keine Normierung)	4.8	1.3	
Textähnlichkeit mit SharedTargetWords $\left(\frac{S}{n_{shared}}\right)$	$\frac{4.8}{4}=1.2$	$\frac{1.3}{1}=1.3$	
Textähnlichkeit mit TotalTargetWords $\left(\frac{S}{n_A+n_B}\right)$	$\frac{4.8}{20+30}\simeq 0.1$	$\frac{1.3}{30+40}\simeq 0.02$	

In diesem Kapitel wird gezeigt, wie sich diese neue Variante auf die Kartenberechnung auswirkt. Vergleichshalber wird zudem eine Variante gezeigt, bei der die Ähnlichkeitswerte gar nicht normiert, sondern nur aufsummiert werden.

5.3 Vorgehen

Es werden Itemsets aus verschiedenen Versuchsreihen verwendet, alle stammen jedoch aus dem Projekt edulap (s. Kap. 23, Projekt edulap) – es handelt sich um Beschreibungstexte von psychologischen Lehrressourcen. Die gefundenen behoften Zielwörter werden nach den beiden Normierungsvarianten in zwei Dreiecksmatrizen geschrieben und die NMDS gerechnet, zum Teil noch prokrustet. Zusätzlich werden die unnormierten Textähnlichkeitswerte berechnet. Da der psychologische Fachbereich der Texte in den Metadaten vermerkt ist, können die Items in der Karte entsprechend eingefärbt werden.

Die drei Rechenverfahren nochmals im Überblick:

- SharedTargetWords
- TotalTargetWords
- keine Normierung (Aufsummierung)

5.4 Resultate

5.4.1 Itemset 1

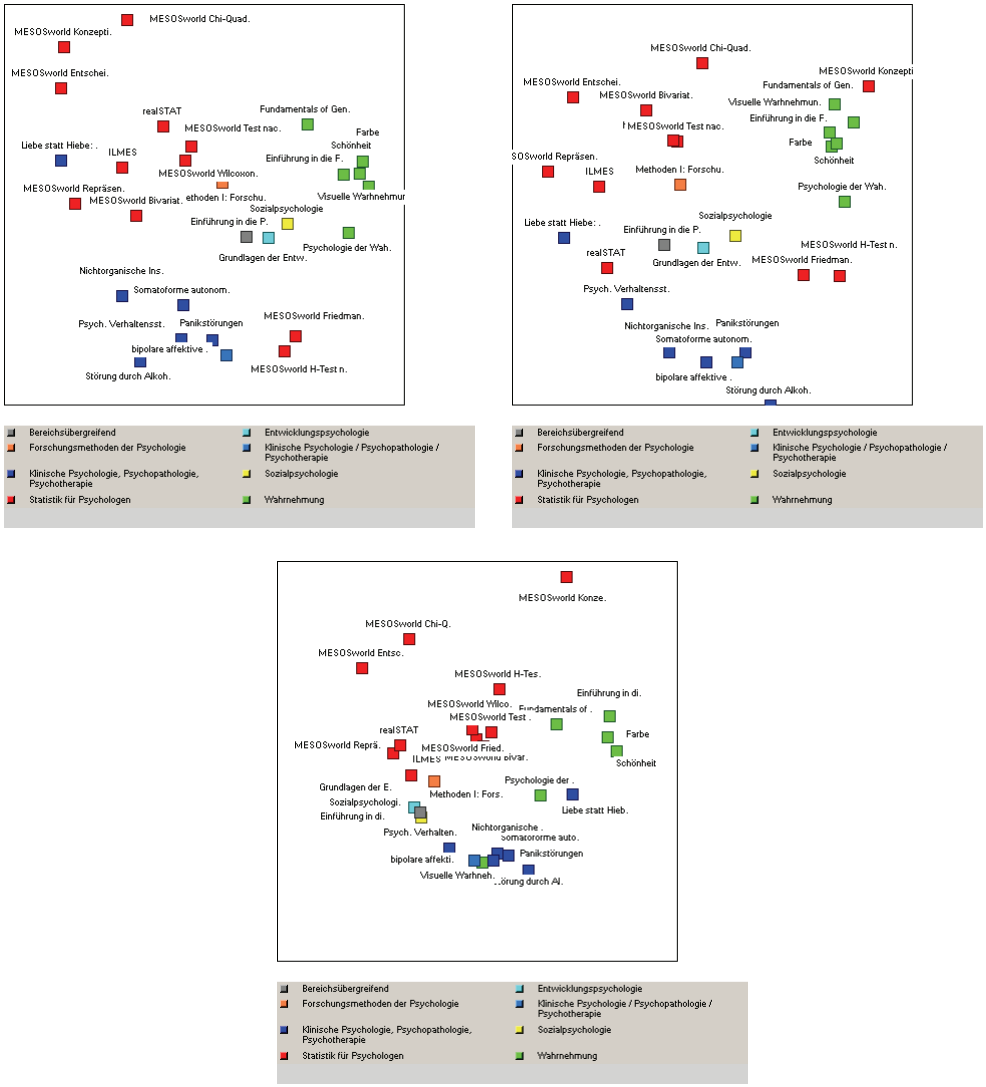


Abbildung 3: Itemset 1. Links oben Normierung nach SharedTargetWords, rechts oben Normierung nach TotalTargetWords; unten Mitte keine Normierung

Beim Itemset 1 (Abb. 3) schneidet interessanterweise «keine Normierung» am besten ab. Nur das Item «Visuelle Wahrnehmung» befindet sich nicht in seinem Cluster, jedoch fällt der Beschreibungstext mit 27 Wörtern nur sehr kurz aus, was eine genauere Lokalisation schwierig macht. Bei Shared- und TotalTargetWords ist das Item «Hiebe statt Liebe» und zwei Statistik-Items aus Mesos nicht im eigenen

Cluster. Bei TotalTargetWords springt zudem noch ein zusätzliches Mesos-Item zum Wahrnehmungsbereich. Insgesamt aber sind sich die beiden Shared- und TotalTargetWords-Karten sehr ähnlich.

Tabelle 4 zeigt die Auswirkung der Normierungen im Vergleich. Ein Ausschnitt der Dreiecksmatrix des Itemsets 1 ist abgebildet. Pro Vergleichspaar sind die drei verschiedenen Ähnlichkeitswerte angegeben (SharedTargetWords/TotalTargetWords/Aufsummierung). Die verschiedenen Bandbreiten der Ähnlichkeitswerte werden deutlich: SharedTargetWords schwankt um den Wert 1, TotalTargetWords zwischen 0 und 1 und die Aufsummierung ist nach oben offen.

Tabelle 4: Ähnlichkeitswerte im Vergleich: SharedTargetWords/TotalTargetWords/Aufsummierung

	Einführung in die Psychologie	**Methoden I: Forsch.+ Stat. I**	**Sozialpsy.**	**Grundlagen der Entw.psy. II**	**ILMES**	**realSTAT**
Einführung in die Psychologie						
Methoden I: Forsch.+ Stat. I	1.25/0.73/60.4					
Sozialpsy.	1.13/0.32/23.46	1.04/0.25/16.47				
Grundlagen der Entw.psy.	1.13/0.37/33.77	1.16/0.2/16.12	1.2/0.23/16.53			
ILMES	1/0.15/8	1.02/0.32/15.25	1.47/0.15/5.87	1.21/0.09/4.85		
realSTAT	1/0.11/6	1.03/0.22/10.36	1.59/0.08/3.18	1/0.04/2	0.93/0.16/2.8	

5.4.2 Itemset 2

Grundsätzlich sind sich die drei Karten auch beim Itemset 2 sehr ähnlich (Abb. 4). Sie alle bilden einen zusammenhängenden Entwicklungscluster ab. Einzig das Item «Podcast Entwicklungspsychologie» springt bei TotalTargetWords, bleibt aber im Entwicklungscluster.

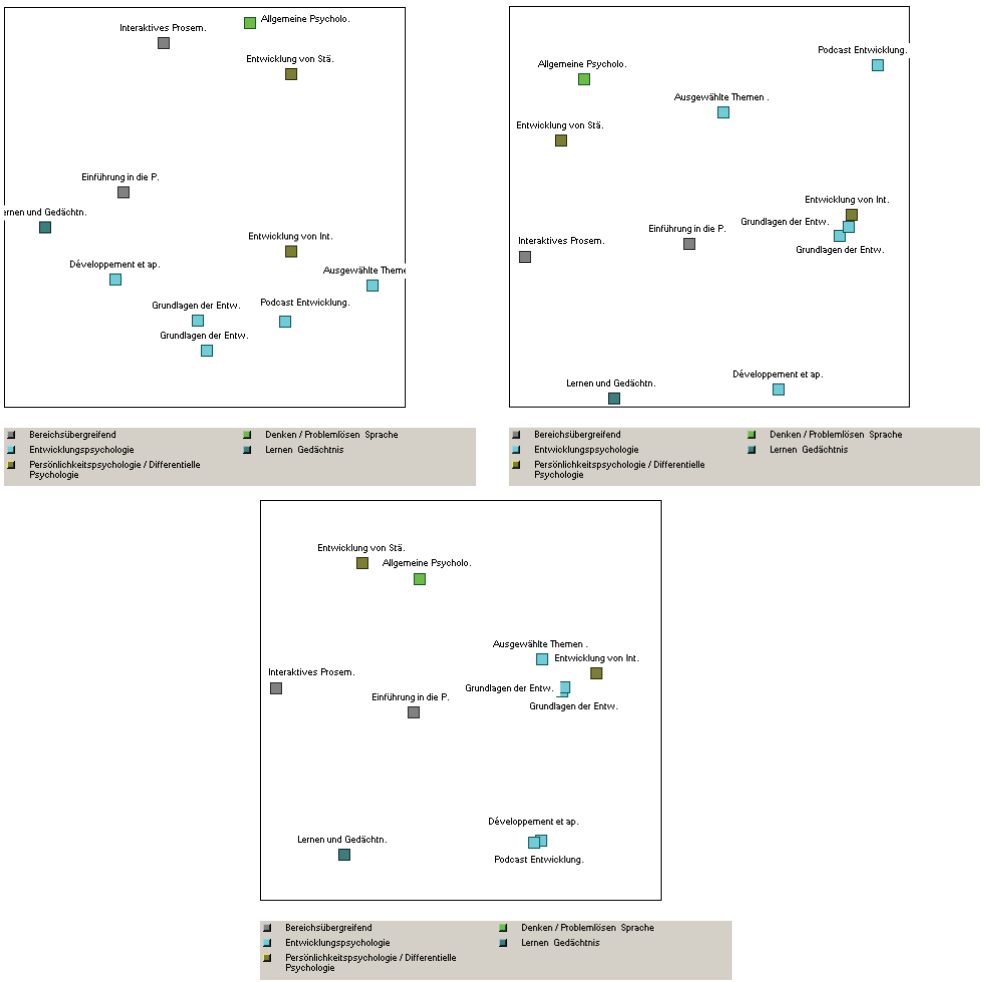


Abbildung 4: Itemset 2. Links oben Normierung nach SharedTargetWords, rechts oben Normierung nach TotalTargetWords; unten keine Normierung

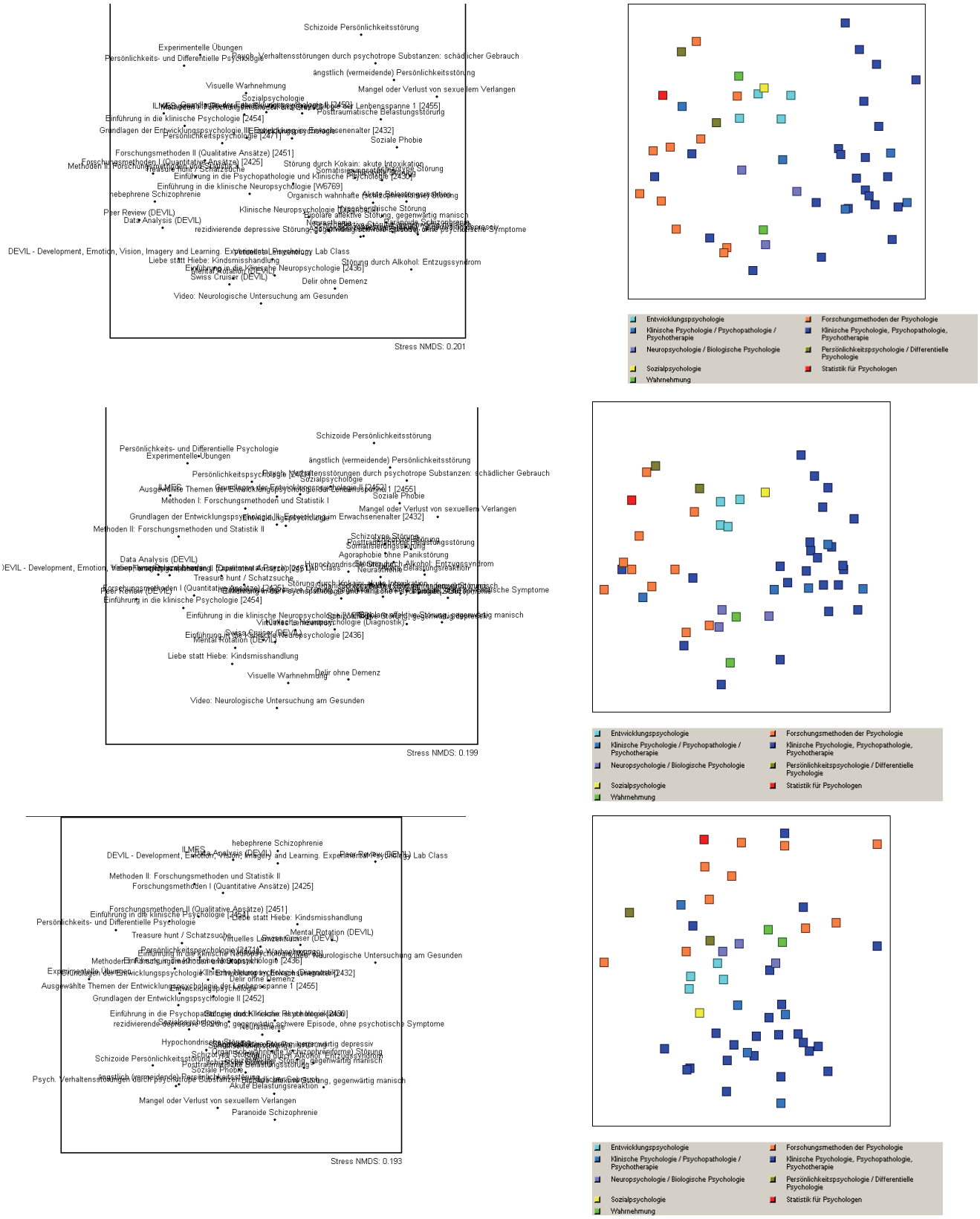


Abbildung 5: Itemset 3. Oben SharedTargetWords; Mitte TotalTargetWords; unten keine Normierung.

5.4.3 Itemset 3

Die Unterschiede in den Karten der drei Normierungsvarianten sind auch beim Itemset 3 nicht gross (Abb. 5). Die Prokrustes-Transformation zwischen Shared- und TotalTargetWords (s. Abb. 6) zeigt nur als Springer den bekannten,kurzen Text «Visuelle Wahrnehmung», der bei TotalTargetWords besser platziert ist. Die übrigen Verschiebungen sind minimal, tendenziell jedoch macht die Karte bei TotalTargetWords einen stimmigeren Eindruck. So ist bei TotalTargetWords die «Soziale Phobie» näher bei Sozialpsychologie, die «Experimentellen Übungen» näher bei den übrigen Methoden und «Einführung in die Neuropsy.» ist näher bei den beiden anderen Neuro-Items.

Abbildung 6: schwarz: SharedTargetWords, rot: TotalTargetWords. Türkis markiert sind die im Text besprochenen Items.

Für alle drei gilt: Die «hebephrene Schizophrenie» liegt mitten im Statistik-Cluster, was inhaltlich falsch ist. «Treasure Hunt» ist ein methodisches Computerspiel, das in der klinischen Diagnostik benutzt wird. Seine Platzierung beim Statistik-Cluster ist somit korrekt. «Liebe statt Hiebe» ist ein Exot, dessen Platzierung zwischen Methoden und Klinischer Psychologie plausibel ist. «Einf. in die klinische P.» ist ausserhalb der klinischen Items, was nicht nachvollziehbar ist. Immerhin stimmt die Verortung am besten bei TotalTargetWords.

5.5 Diskussion

Deutliche Unterschiede sind zwischen den Normierungsvarianten nicht auszumachen. Mal schneidet SharedTargetWords besser ab, mal TotalTargetWords, mal gar keine Normierung. Da die Herleitung der Normierung von TotalTargetWords aber am plausibelsten ist, ist dieser Normierung den Vorzug zu geben.

6 Hofgewichtung

6.1 Überblick

Um die geringe Menge an Semantik in den Höfen zu verstärken, wird die Hofähnlichkeitsberechnung um einen Faktor ergänzt, der den errechneten Ähnlichkeitswert multipliziert. Diese als «Signalverstärkung» gedachte Massnahme hat jedoch überraschend wenig Einfluss auf die Strukturierung der semantischen Karten.

6.2 Einleitung

Die Ähnlichkeit zweier Texte wird durch die Hofähnlichkeiten bestimmt. Der Hofähnlichkeitswert setzt sich aus dem Wert der beiden Zielwörter und den aufsummierten Werten der Hofwörter zusammen. Zwei Texte haben also schon eine gewisse Ähnlichkeit, wenn zwei ähnliche Zielwörter darin gefunden werden. Haben diese Zielwörter auch noch ähnliche Hofwörter, dann steigt diese Ähnlichkeit.

Allerdings ist die Information, die in den Höfen steckt, spärlich: Nur selten werden ähnliche Wörter in den Höfen gefunden. Welche Auswirkung hat es, wenn diese Information stärker gewichtet wird, als durch die bisher verwendete einfache Cosinusfunktion? Lassen sich dadurch qualitative Verbesserungen erzielen?

Die Formel der Ähnlichkeitsberechnung zweier Hofwörter (V) wird also um einen Gewichtungsfaktor (g) erweitert:

$$V = \cos(d \cdot \frac{90}{Hof + 1}) \cdot g$$

Dieser Faktor kommt bei der Ähnlichkeitsberechnung der Hofwörter zum Einsatz, nicht aber bei den Zielwörtern selbst.

Als Datensatz verwenden wir eine Auswahl der edulap-Daten (s. Kap. 23, Projekt edulap), sowie der Daten der Stadtdebatte (s. Anhang 2.1, Beschreibung des Webforums Stadtdebatte).

6.3 Vorgehen

Die beiden Datensätze aus den Projekten «edulap» und «Stadtdebatte» werden behoft und die Hofähnlichkeiten jeweils mit verschiedenen Hofgewichtungsfaktoren berechnet und nach TotalTargetWords (s. Kap. 5, Normierung der Textähnlichkeitswerte: SharedTargetWords vs. TotalTargetWords) normiert. Anschliessend wird die Strukturierung der resultierenden Karten untersucht.

6.4 Resultate

6.4.1 Datensatz edulap

Nachfolgende Tabelle zeigt exemplarisch, wie sich das Hofgewicht (Verstärkungsfaktor) auf die DEM auswirkt. Für jeden Paarwert wird die originale Ähnlichkeit (Hofgewicht 1) und die gewichteten Ähnlichkeiten (Hofgewicht 10 und 100) angegeben.

Tabelle 5: Ausschnitt aus der Dreiecksmatrix. Die Textähnlichkeiten wurden mit TotalTargetWords normiert. Angegeben sind die paarweisen Ähnlichkeiten mit einer Gewichtung von 1, 10 und 100.

	Methoden I: Forschungsmethoden und Statistik I	**Methoden II: Forschungsmethoden und Statistik II**	**Sozialpsychologie**
Methoden I: Forschungsmethoden und Statistik I			
Methoden II: Forschungsmethoden und Statistik II	0.97 / 1.73 / 9.38		
Sozialpsychologie	0.25 / 0.4 / 1.9	0.1 / 0.1 / 0.1	
Persönlichkeits- und Differentielle Psychologie	0.02 / 0.02 / 0.02	0.08 / 0.08 / 0.08	0.03 / 0.03 / 0.03

Skaliert man die Werte, ergeben sich nur geringfügige Veränderungen (s. Abb. 7). Der Stresswert ist bei allen Varianten etwa 0.2. Auch die Prokrustes-Transformation der Originalkarte mit der 10er-Karte zeigt keine strukturellen Verbesserungen (Abb. 8).

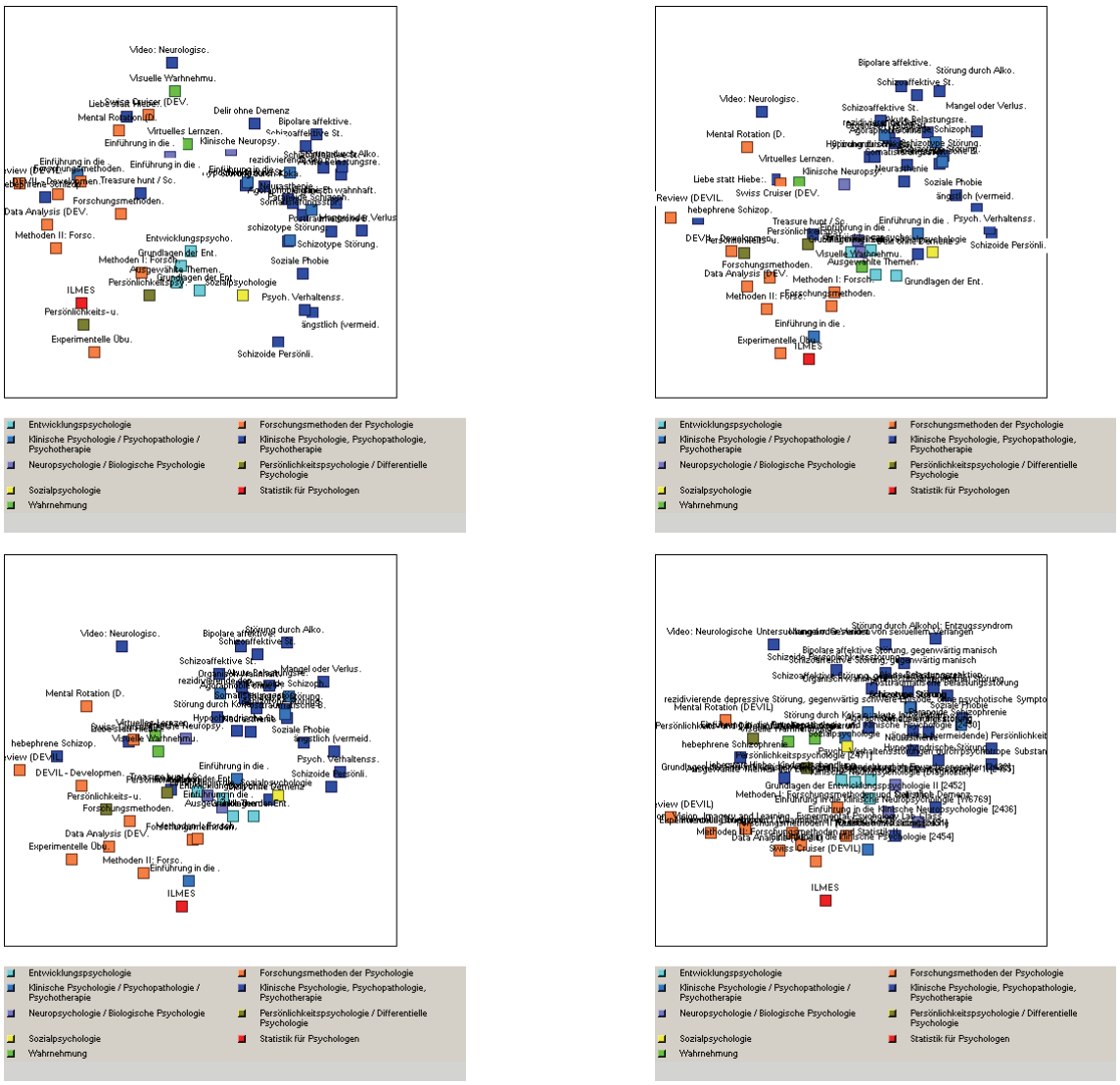


Abbildung 7: Veränderung der Struktur durch Vergrössern des Hofgewichtes. Oben links: Hofgewicht 1; oben rechts: Hofgewicht 2; unten links: Hofgewicht 10; unten rechts: Hofgewicht 100

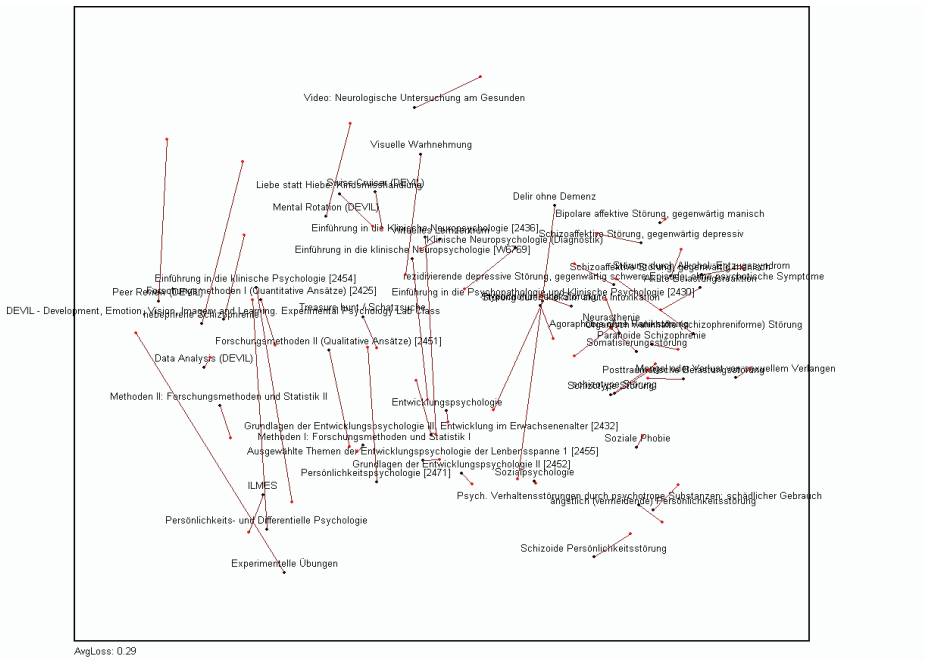


Abbildung 8: Prokrustes, Hofgewicht 1 (schwarz) auf Hofgewicht 10 (rot)

6.4.2 Datensatz Stadtdebatte

Es wurde dieselbe Auswahl von Texten genommen wie bei Kapitel 28.4.2 (Strukturierung mehrerer Foren). Wieder wurden die Hofähnlichkeiten nach TargetWords normiert. Das Hofgewicht wurde einmal auf 1 und einmal auf 100 gesetzt. Abbildung 9 zeigt die beiden resultierenden Karten, sowie deren Prokrustes-Transformation. Es sind keine qualitativen Veränderungen festzustellen. Die wenigen Verschiebungen sind nicht von denjenigen zu unterscheiden, die von der teilweise zufälligen Auswahl der TargetWords herrühren.

6.5 Diskussion

Die Hofgewichtung hat einen überraschend geringen Einfluss auf die Skalierung der Hofähnlichkeitswerte. Wenn zufällige Faktoren ins Spiel kommen – wie das bei der Auswahl der TargetWords der Fall ist – sind die Einflüsse der Hofgewichtung nicht mehr davon zu unterscheiden.

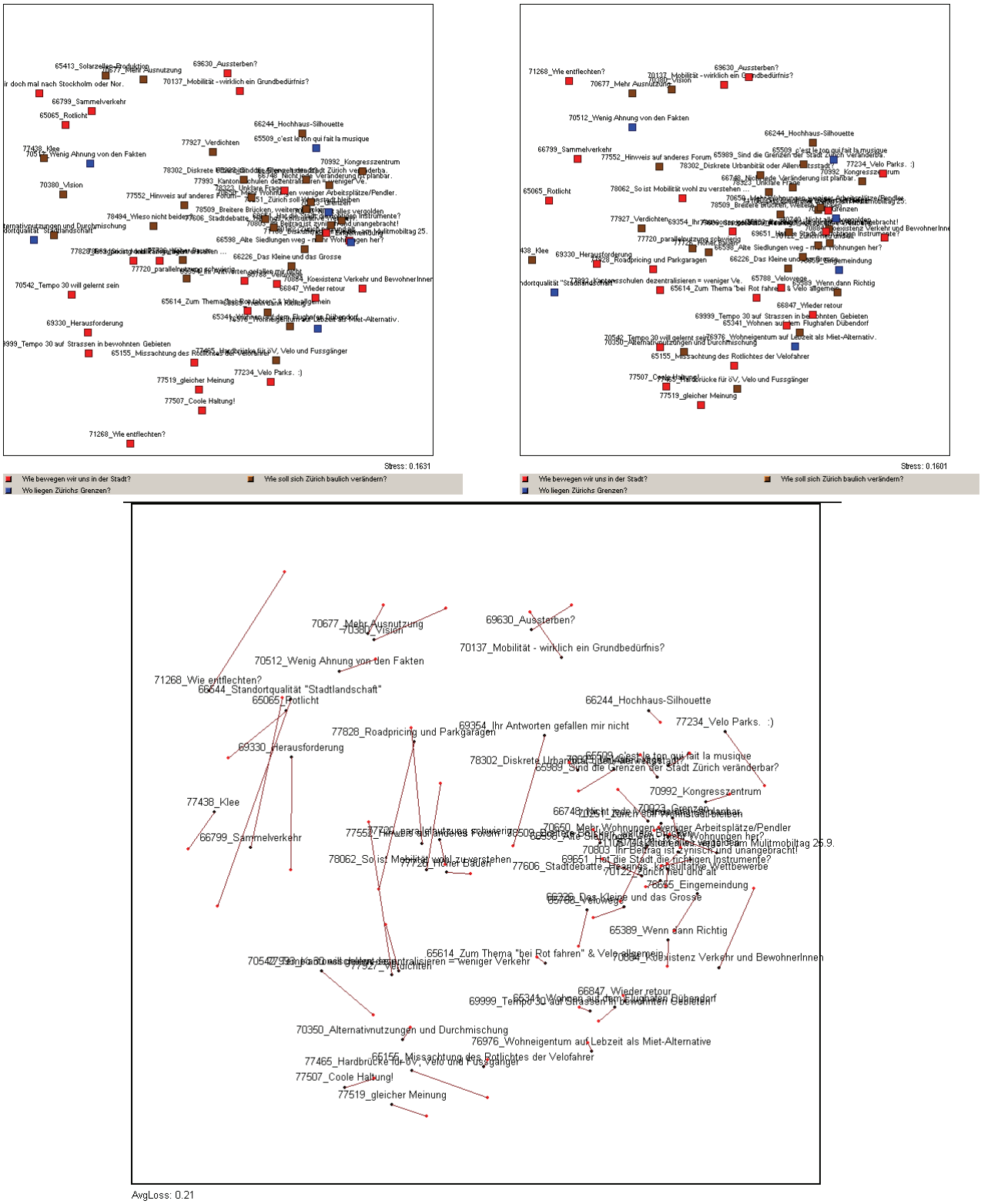


Abbildung 9: Oben links: Hofgewicht 1; oben rechts: Hofgewicht 100; unten Mitte: Prokrustes-Transformation der beiden Karten. Es sind keine nennenswerten Verschiebungen erkennbar.

7 TargetWords im Hof

7.1 Überblick

Werden Texte aufgrund eines einzigen TargetWords verglichen, treten Scheinähnlichkeiten auf, wenn das TargetWord auch im Hof vorkommt. Der Wert des Hofwortes muss deshalb auf Null gesetzt werden.

7.2 Einleitung

Wenn Textvergleiche anhand eines einzigen TargetWords durchgeführt werden (wie in Kap. 24, Wikipedia-Experiment), entstehen falsche Hofähnlichkeiten, wenn das TargetWord auch im Hof vorkommt. Grund dafür ist die zusätzliche Aufsummierung des Hofwortes. Wenn das TargetWord, nehmen wir «Luft», nicht nur im Zentrum des Hofes vorkommt, sondern auch im Hof selbst, resultieren daraus Ähnlichkeiten mit allen anderen Höfen, die das Wort ja im Zentrum haben. Dies hat zur Folge, dass eine Scheinähnlichkeit produziert wird: Nicht die inhaltliche Verwendung des Wortes wird wiedergegeben, sondern das numerische Auftreten des Stichwortes. Das ist nicht erwünscht und muss verhindert werden.

7.3 Vorgehen

Die Lösung ist so simpel wie effektiv: Tritt das Zielwort im Hof auf, wird sein Hofwert (der den Abstand zum Zielwort repräsentiert) auf Null gesetzt. Dadurch geht sein Einfluss auf die Ähnlichkeitsberechnung zweier Höfe verloren.

	Id	Kategorie	Wi	Ur:	Assoziation	Value
	3	Luft-Luft-Rakete	luft	luft	welcher	0.258819045102521
	3	Luft-Luft-Rakete	luft	luft	kampfmittel	0.5
	3	Luft-Luft-Rakete	luft	luft	luftkampf	0.707106781186548
	3	Luft-Luft-Rakete	luft	luft	eingesetzt	0.866025403784439
	3	Luft-Luft-Rakete	luft	luft	anfangsphase	0.965925826289068
	3	Luft-Luft-Rakete	luft	luft	luft	1
▶	3	Luft-Luft-Rakete	luft	luft	luft	0
	3	Luft-Luft-Rakete	luft	luft	raketen	0.866025403784439
	3	Luft-Luft-Rakete	luft	luft	ersten	0.707106781186548
	3	Luft-Luft-Rakete	luft	luft	weltkrieg	0.5
	3	Luft-Luft-Rakete	luft	luft	zweiten	0.258819045102521

Abbildung 10: Tritt das Zielwort im Hof selbst auf (hier schwarz markiert), wird sein Wert auf Null gesetzt. Somit kann es nicht mehr störend in die Berechnung der Ähnlichkeit zweier Höfe einwirken.

7.4 Resultate

Die Auswirkungen sind beträchtlich. Nachfolgend sind zwei Karten abgebildet. Beide stammen aus dem erwähnten Wikipedia-Experiment. Der Wikipedia-Index wurde auf das Stichwort «Luft» hin durchsucht. 58 resultierende Einträge wurden dann nach diesem Stichwort behoft. Links ist die Karte abgebildet, bei der die Wörter «Luft» im Hof mitgerechnet wurden, rechts diejenige ohne.

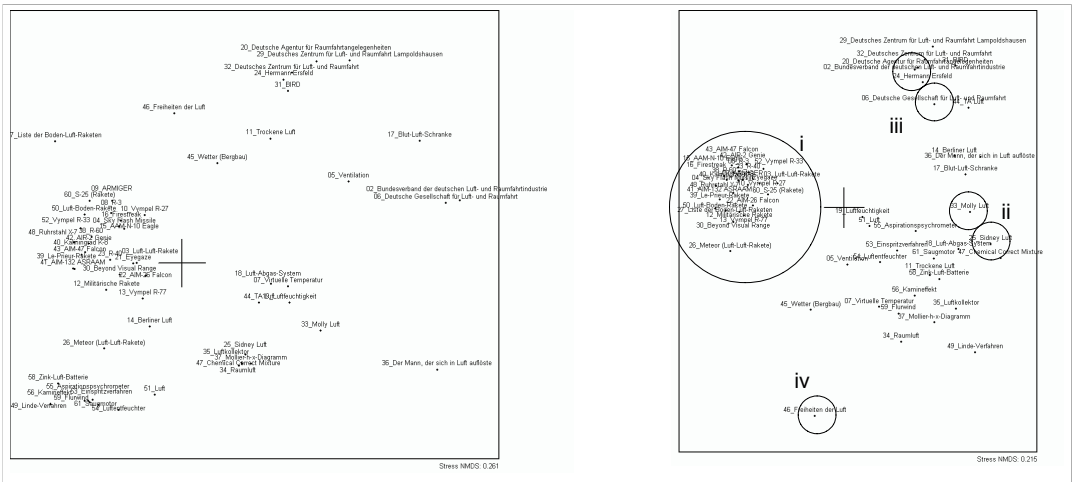


Abbildung 11: Links wurde das wiederholte Vorkommen des Zielwortes im Hof mitgerechnet, rechts wurde dessen Wert auf Null gesetzt, damit Scheinähnlichkeiten vermieden werden (die eingekreisten Bereiche werden im Text erläutert).

In der Grundstruktur sind sich die Karten ähnlich, jedoch sind rechts einige Items besser eingeordnet:

i. Der Cluster mit den Raketen ist schärfer abgegrenzt.
ii. Die beiden Personen «Molly Luft» und «Sydney Luft» sind nicht mehr mitten im «Luft-als-Stoff»-Cluster.
iii. Die Items 2 (Bundesverband der deutschen Luft- und Raumfahrtindustrie) und 6 (Deutsche Gesellschaft für Luft- und Raumfahrt) sind nicht mehr separat, sondern im passenden Luft- und Raumfahrtcluster
iv. Das Item 46 (Freiheiten der Luft) – es geht um Luftrechte im kommerziellen Luftverkehr – ist hingegen schlechter platziert.
v. Es gibt keine falschen Cluster mehr.

7.5 Diskussion

Durch die Null-Setzung von TargetWords im Hof wird die Scheinähnlichkeit des numerischen Auftretens eines Begriffes eliminiert. Das Vorgehen ist unumgänglich, wenn Texte aufgrund eines einzigen TargetWords verglichen werden und ist eine pragmatische Lösung für das Problem der verzerrten Ähnlichkeit.

8 Vergleich Stoppwörter: Urliste vs. Snowball

8.1 Überblick

Texte müssen mittels Stoppwortlisten entrauscht werden, bevor die Behofung stattfinden kann. Die Urliste ist eine statistisch zusammengestellte Liste, während Snowball auf linguistischen Kriterien beruht. Beide produzieren ähnliche Resultate, jedoch komprimiert Snowball – bedingt durch ihren dreifachen Umfang – die Texte stärker, was die spätere Verarbeitung beschleunigt.

8.2 Einleitung

In natürlich-sprachlichen Texten ist viel Rauschen in Form von Füll- und Strukturwörtern. Die Hofmethode begegnet dem, indem sie sog. Stoppwörter aus dem Text entfernt. Der Einfachheit halber wurden diese Wörter bisher als die 100 häufigsten[15] der zugrunde liegenden Sprache definiert. Die Idee dahinter ist, dass allzu häufig verwendete Wörter kein semantisches Gehalt mehr besitzen. Diese Liste wurde noch bearbeitet: Zum Teil waren in der deutschen Liste Wörter vorhanden, denen doch eine gewisse Semantik zugeschrieben wurde; während andere Wörter händisch eingefügt wurden. Schliesslich umfasste die Liste 88 Wörter. Weil das die erste Liste war, mit der wir arbeiteten, nennen wir sie «Urliste».

Die Stoppwörterliste wird im klassischen Information Retrieval aber aufgrund linguistischen Gesichtspunkten zusammengestellt. Aus dem Snowball-Projekt[16] gingen für verschiedene Sprachen solche Listen hervor; für Deutsch umfasst sie 231 Wörter[17]. Sie ist somit drei Mal länger als die Urliste und könnte das Rauschen einerseits stärker reduzieren, andererseits auch mehr Information vernichten. Zudem steigt der Rechenaufwand zum Entrauschen der Texte, dafür sind die Texte danach stärker komprimiert, was die weitere Verarbeitung beschleunigt.

Eine andere Möglichkeit wäre es, die Stoppwörter aus dem aktuellen Textkorpus zu generieren, beispielsweise die 60 häufigsten Wörter zu wählen. Der Vorteil dieses Vorgehens wäre auch sein Nachteil: Die Stoppwörter wären zwar auf das aktuelle Textmaterial zugeschnitten, müssten aber bei grösseren

15 http://wortschatz.uni-leipzig.de/html/wliste.html
16 http://snowball.tartarus.org/texts/introduction.html
17 http://snowball.tartarus.org/algorithms/german/stop.txt

Textmutationen laufend aktualisiert werden. Wir verfolgen diese Idee in vorliegender Arbeit nicht weiter, möchten sie aber nicht unerwähnt lassen, da hierin sicherlich weiteres Optimierungspotenzial steckt.

8.3 Vorgehen

In diesem Kapitel werden die Urliste und die deutschsprachige Snowball-Liste miteinander verglichen, um eine Aussage darüber zu ermöglichen, welche Liste das Rauschen auf geeignetere Weise reduziert. Als Datengrundlage dienen 70 Texte des edulap-Projekts (Kap. 23, Projekt edulap).

8.4 Resultate

Die 70 Testitems wurden entweder mit Hilfe der Urliste oder der Snowball-Liste entrauscht, anschliessend nach der KeywordII-Methode (Kap. 15, KeywordII-Analyse) behoft. Erwartungsgemäss sind die Unterschiede nicht gross (s. Abb. 12).

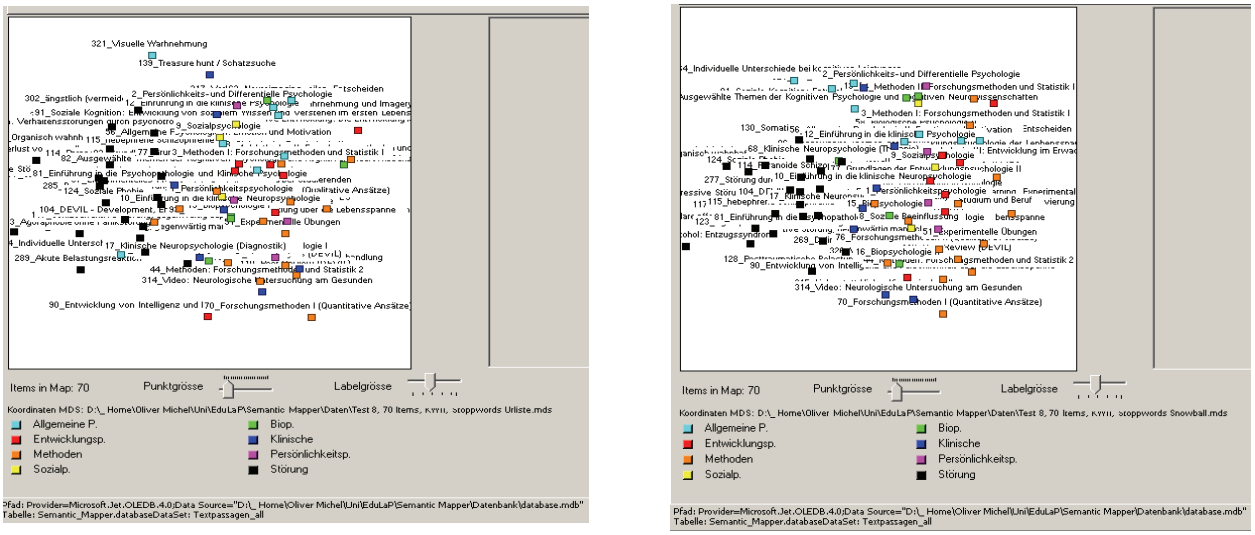


Abbildung 12: Link ist die Karte basierend auf mit der Urliste entrauschten Texten, rechts mit der Snowballliste. Es sind keine grossen Unterschiede sichtbar, jedoch scheint die Snowball-Variante etwas deutlicher zu gruppieren.

Als Gütekriterium dient die Gruppierung (nach Augenschein) der psychologischen Kategorien, die unterschiedlich eingefärbt wurden. In der Snowball-Karte erscheinen die Texte etwas «gruppierter»; die Items der einzelnen Kategorien liegen in geringem Masse näher beisammen.

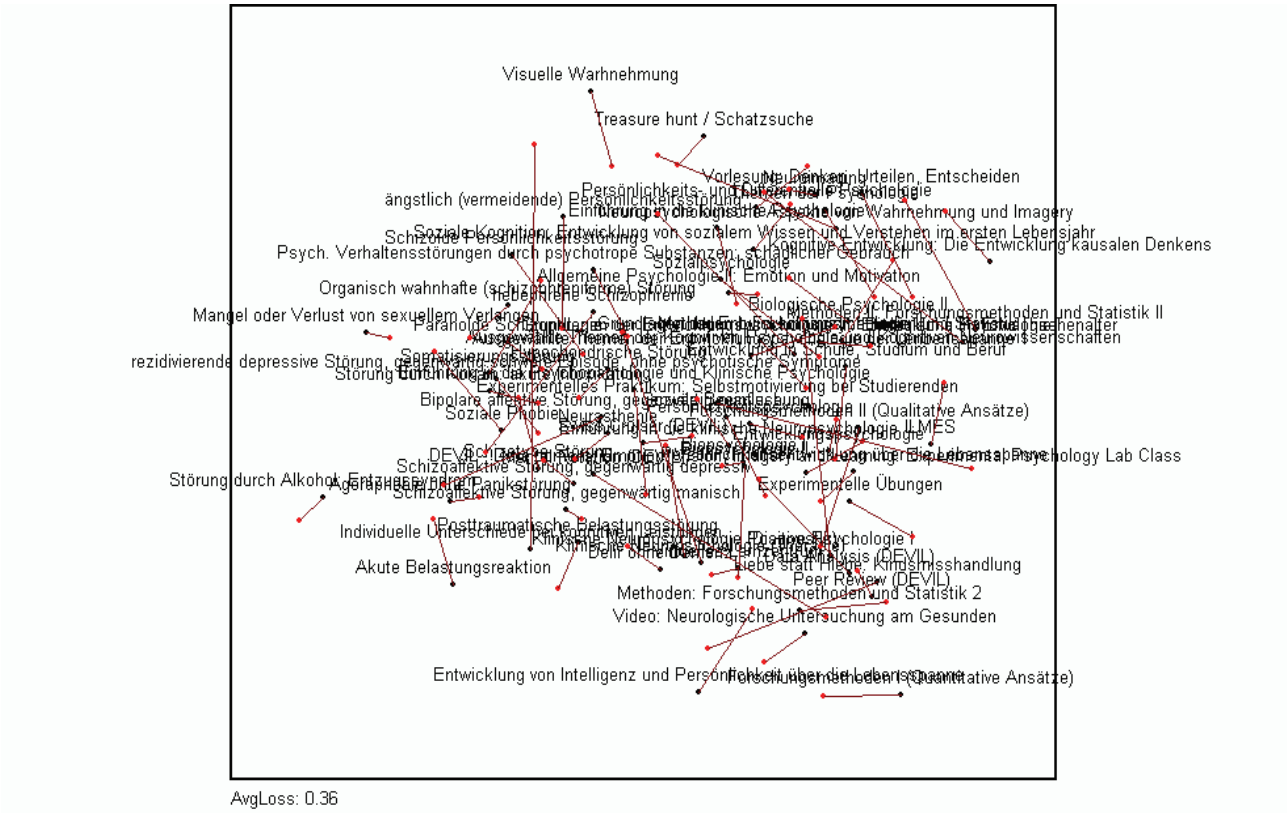


Abbildung 13: Prokrustes-Transformation der beiden Stoppwords-Varianten. Die schwarzen Punkte stammen von der Urlisten-Berechnung, die roten von Snowball. Deutlichster Unterschied: Das Item «Individuelle Unterschiede bei kognitiven Leistungen» springt bei der Snowball-Variante in seine zugehörige Gruppe.

Die Prokrustes-Transformation bestätigt diesen Eindruck, zudem wird ein deutlicher Fehler der Urlisten-Karte sichtbar: Das Item «Individuelle Unterschiede bei kognitiven Leistungen» springt bei der Snowball-Karte zu seiner Gruppe, während es bei der Urlisten-Karte abgeschieden unterhalb des Störungsbereichs weilt (s. Abb. 13).

8.5 Diskussion

Das Resultat ist mager: Ein einziges Item, das bei der Urlisten-Variante offensichtlich falsch zu liegen kam, wurde mit der Snowball-Variante besser eingeordnet, zudem ist die restliche Anordnung mit Snowball etwas gefälliger.

Die Entscheidung für oder gegen eine der Listen ist weniger abhängig von diesem konkreten Resultat, als von theoretischen Überlegungen: Snowball wurde aufgrund linguistischen Gesichtspunkten zusammengestellt, während die Urliste statistisch begründet ist. Da die Hofmethode als solche ebenfalls statistisch funktioniert, wäre die Urliste passender, da sie keine Top-Down-Linguistik benötigt. Eine aus dem aktuellen Datenpool generierte Liste wäre in Betracht zu ziehen, jedoch existieren hierzu keine Untersuchungen.

Beim edulap-Projekt geben wir der Snowball-Liste den Vorzug, da sie die Texte stärker komprimiert und somit die weitere Verarbeitung beschleunigt. Durch den grösseren Umfang ist sie zudem robuster, was die Aufnahme neuer Texte betrifft.

9 Rauschreduktion mit und ohne Stemming

9.1 Überblick

Das Stemming führt Wörter auf ihren grammatikalischen Wortstamm zurück. Dadurch wird das natürlich-sprachliche Rauschen beim Hofvergleich reduziert. Stemming ist somit eine Alternative zum unscharfen Wortvergleich. Der Effekt ist allerdings marginal und lohnt den Aufwand nicht.

9.2 Einleitung

Das Rauschen in den natürlich-sprachlichen Texten ist sehr hoch: Wörter werden gebeugt und flektiert, Synonyme und Homonyme verwendet, die gleichen Dinge verschiedentlich umschrieben und dergleichen mehr. Eine Möglichkeit der Rauschunterdrückung ist das *Stemming*: Wörter werden maschinell auf ihren grammatikalischen Wortstamm reduziert. In der klassischen Computerlinguistik ist das eine essentielle Methode, um gleiche Begriffe (aber mit unterschiedlichen Wortformen, weil beispielsweise flektiert) zu identifizieren. Die Kosten eines Stemmings sind jedoch hoch: Die Wörter müssen mit aufwändigen Algorithmen auf den Wortstamm gekürzt werden, die Implementierung ist fehleranfällig und der Stemming-Vorgang braucht sehr viel Rechenaufwand. Die Hofmethode verfolgt einen gänzlich anderen Ansatz, indem sie eine unscharfe Ähnlichkeit (s. Kap. 2.1.5, Unscharfe Ähnlichkeit) zwischen Wörtern berechnet. Dies ist ein sehr pragmatischer Ansatz. In diesem Kapitel wird anhand des edulap-Projekts (s. Kap. 23, Projekt edulap) untersucht, ob das Stemming eine geeignetere Methode der Rauschreduktion ist.

9.3 Vorgehen

Als Sample wurden 70 Items aus dem edulap-Projekt genommen. Die Texte wurden mit zwei unterschiedliche Varianten behandelt:

Variante A

- einfache Entrauschung durch Entfernen der Stoppwörter

Variante B

- einfache Entrauschung durch Entfernen der Stoppwörter, sowie zusätzliches Stemming

Als Stemmer wurde eine quelloffene Bibliothek verwendet.[18] Bei beiden Varianten wurden die Keywords anhand der KeywordII-Methode (s. Kap. 15, KeywordII-Analyse) errechnet. Bei der Variante A wurden die Texte anschliessend nach einer unscharfen Ähnlichkeit zu den Keywords durchsucht, bei der Variante B mussten die Keywords genau übereinstimmen (weil ansonsten – so die Überlegung – die Entrauschung überhand nehmen würde). Dies wird in Unterkapitel 9.4.1 überprüft.

Nach der Behofung blieben 69 Lernitems übrig, da ein Beschreibungstext zu kurz war, als dass die KeywordII-Methode fündig geworden wäre. Den Items wurde manuell eine von drei Kategorien zugeteilt:

- Vorlesung mit Inhalt Methoden oder Statistik
- Vorlesung mit anderem Inhalt
- Störungsbild

Diese Kategorien ergaben sich empirisch aus der Expertenbefragung, die im Zusammenhang mit der Testreihe Expertenkarten (s. Kap. 23, Projekt edulap) durchgeführt wurde. Die Experten scheinen diese Einteilung zu bevorzugen, sie bildet also ein Gütekriterium für das Stemming.

9.3.1 Beispiel Entrauschungsprozedere

Der Rohtext wird (in Variante B) zwei Verarbeitungsprozedere unterzogen, bevor er behoft wird: Erst findet die Rauschreduzierung durch Aussortierung der Stoppwörter und Satzzeichen statt, danach das Stemming. An einem konkreten Text wird dies demonstriert:

Originaltext:

> Das Modul gibt eine Einführung in die wichtigsten Fragestellungen, Forschungsmethoden, Theorien und Ergebnisse der Persönlichkeitsforschung. Gegenstand ist die Beschreibung und Erklärung reliabler Unterschiede zwischen Personen, deren Messung und Organisation, sowie deren Konsequenzen. Themenbereiche sind unter anderem Temperament, Charakter, Intelligenz, Kreativität und Affektivität. Ferner werden einzelne Persönlichkeitstheorien über das Lesen der Literatur, vertiefende Aufgaben und Austausch im Online-Forum und Diskussion an den Präsenzterminen erarbeitet. Die Veranstaltung hat einen praktischen Teil, in dem wichtige repräsentative Erfassungsmethoden zu Persönlichkeitsmerkmalen im Selbstversuch erprobt werden.£

entrauscht:

> modul gibt einführung wichtigsten fragestellungen forschungsmethoden theorien ergebnisse persönlichkeitsforschung gegenstand beschreibung erklärung reliabler unterschiede zwischen personen deren messung organisation sowie deren konsequenzen themenbereiche anderem temperament charakter

18 http://www.codeproject.com/KB/recipes/destemming.aspx?fid=14644&df=90&mpp=25&noise=3&sort=Position&view=Quick&select=430926

intelligenz kreativität affektivität ferner einzelne persönlichkeitstheorien lesen literatur vertiefende aufgaben austausch online forum diskussion präsenzterminen erarbeitet veranstaltung praktischen teil wichtige repräsentative erfassungsmethoden persönlichkeitsmerkmalen selbstversuch erprobt

Stemmed:

modul gibt einfuhr wichtig fragestell forschungsmethod theori ergebniss personlichkeitsforsch gegenstand beschreib erklar reliabl unterschied zwisch person der messung organisation sowi der konsequenz themenbereich and temperament charakt intelligenz kreativitat affektivitat fern einzeln personlichkeitstheori les literatur vertief aufgab austausch onlin forum diskussion prasenztermin erarbeitet veranstalt praktisch teil wichtig reprasentativ erfassungsmethod personlichkeitsmerkmal selbstversuch erprobt

9.4 Resultate

Abbildung 14 zeigt die gemittelte Expertenkarte: Durch die Mittelung sind die Cluster sehr eng, jedoch sind die schwierigen Items gut erkennbar.

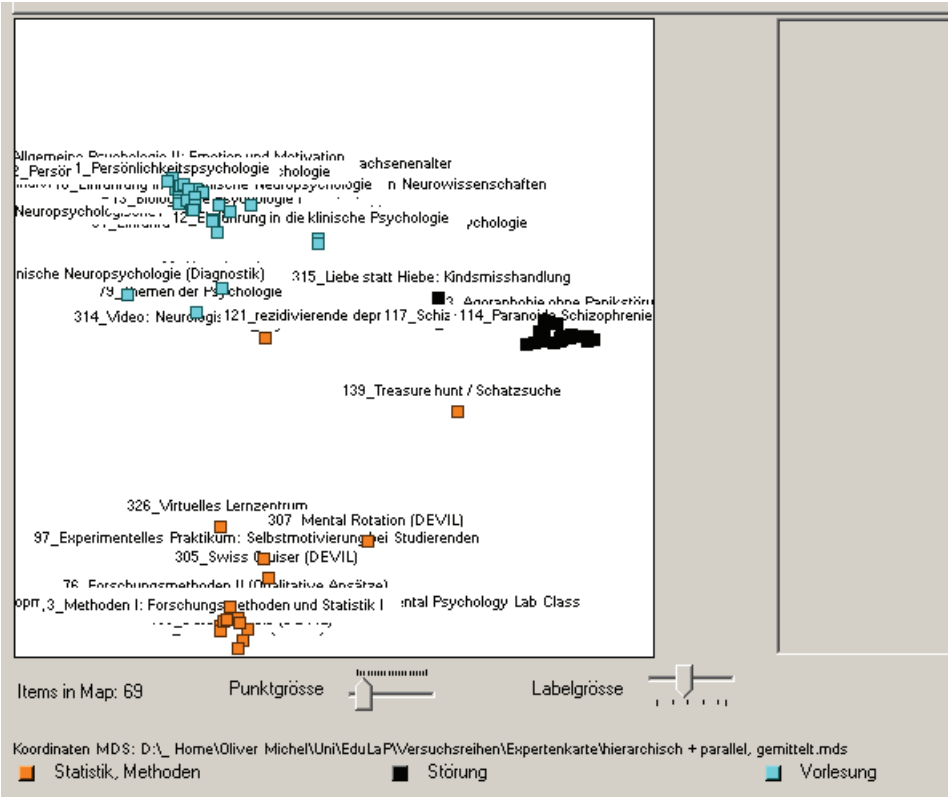


Abbildung 14: gemittelte Expertenkarte, Einfärbung nach manueller Kategorisierung

Abbildungen 15 und 16 zeigen die Strukturierungen durch die HM nach den beiden Varianten.

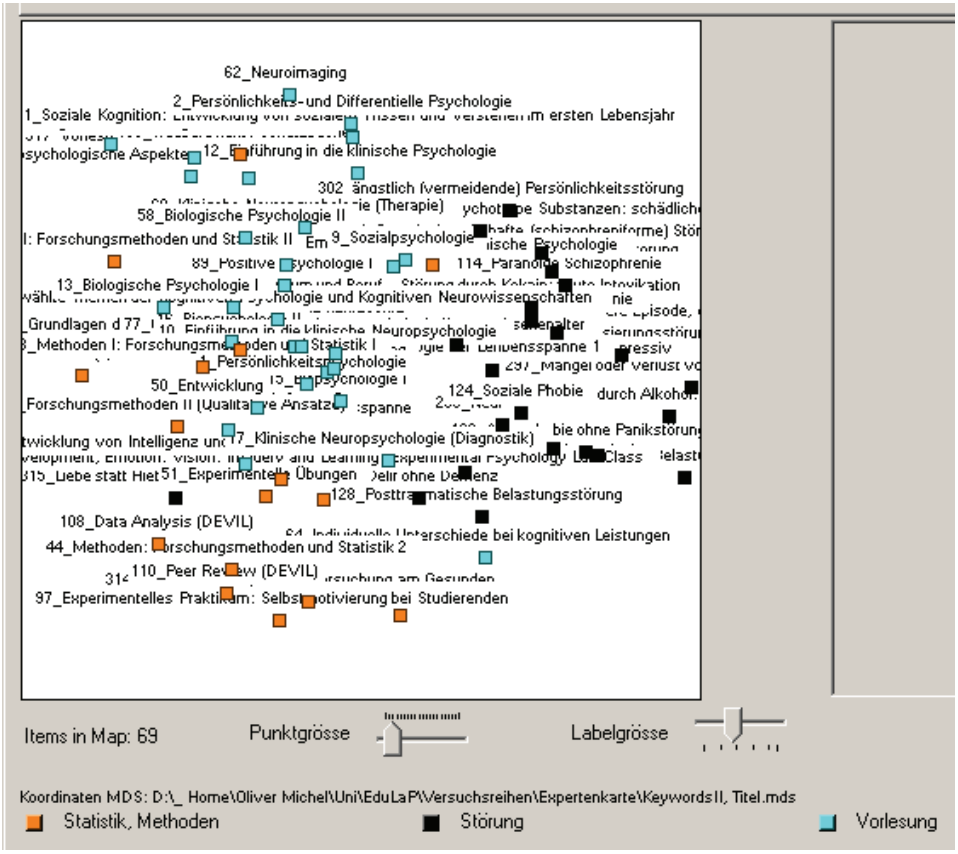


Abbildung 15: Karte der entrauschten Texte ohne Stemming (Variante A)

Ohne Stemming produziert die HM Ähnlichkeiten, die die Makrostruktur der Expertenkarte grob wiedergibt. Die Störungsbilder sind alle – bis auf das schwierige, weil unklare, Item «315_Liebe statt Hiebe» recht klar beisammen. Links unten sind die Statistik-Vorlesungen, die sich gegen oben mit den übrigen Vorlesungen vermischen. Eine Clusterung gibt es nicht, jedoch zeigt die Einfärbung deutlich die Bereiche der vorgängig getätigten Kategorisierung.

Werden die Texte gestemmed (Variante B, Abb. 16) ergeben die errechneten Ähnlichkeiten tendenziell deutlichere Kategorisierungen.

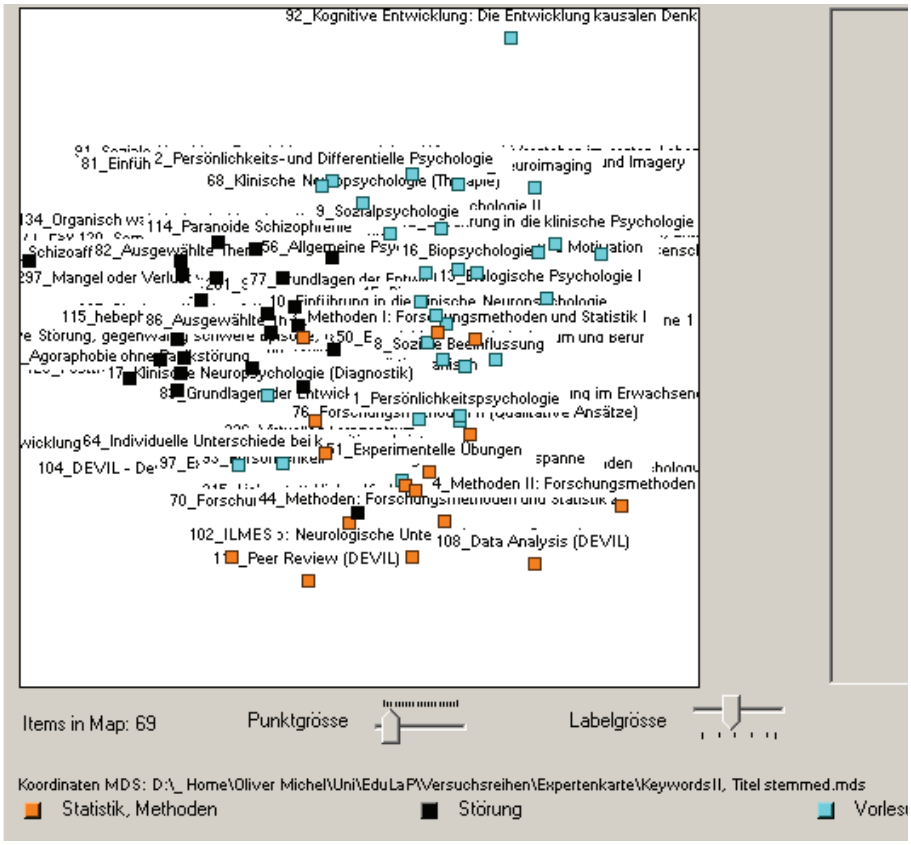


Abbildung 16: Karte der entrauschten Texte mit Stemming (Variante B)

Um die Kartenunterschiede besser zu erfassen, wurden Prokrustes-Transformationen in Abbildung 17 durchgeführt. Der Average Loss ist sehr ähnlich, ob mit oder ohne Stemming. Somit ist der Effekt auf der Ebene der Makrostruktur eher marginal. Die Intraclusterstruktur scheint deutlichere Abweichungen zu haben, jedoch wurden diese hier nicht weiter untersucht.

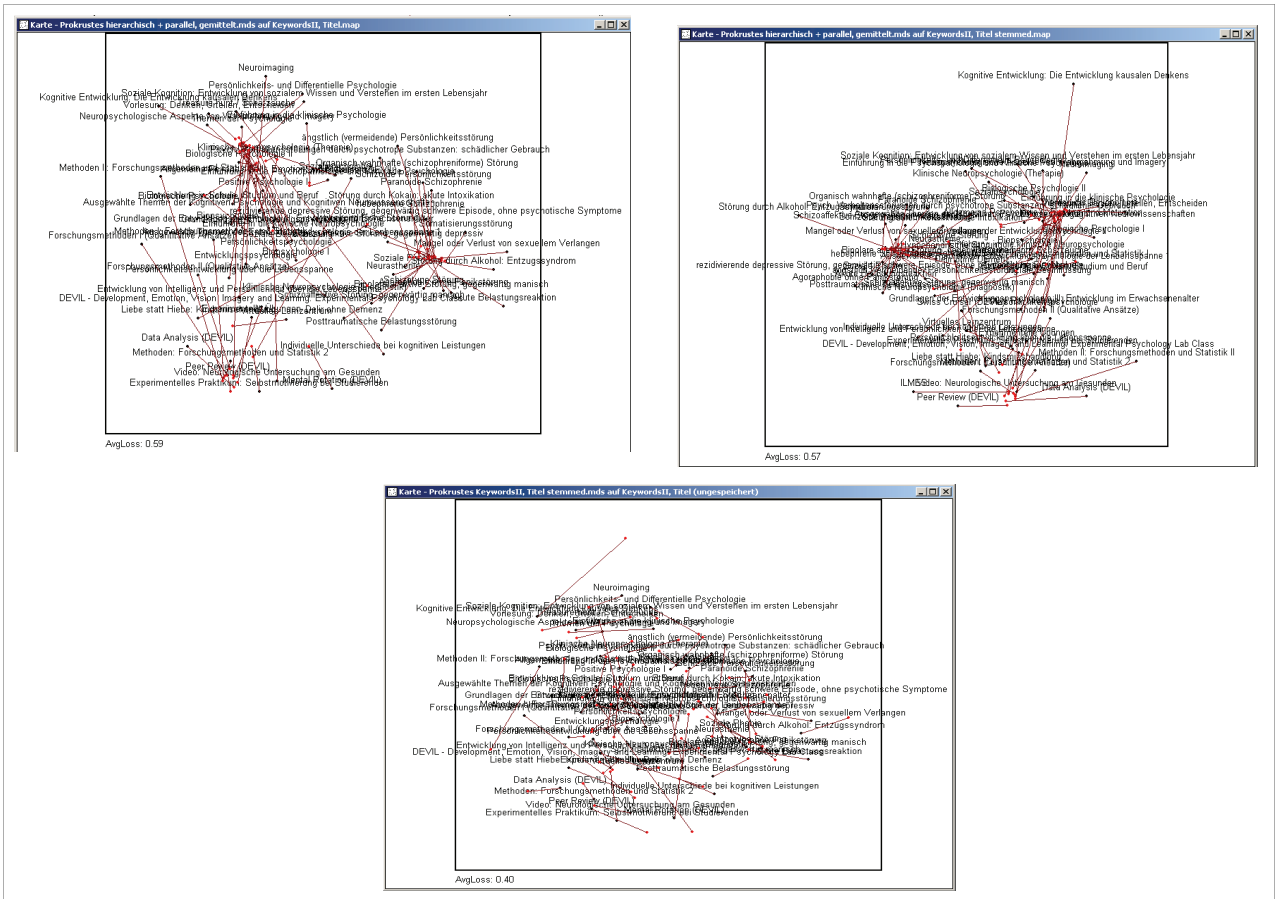


Abbildung 17: Prokrustes-Transformationen der entrauschten Texte mit der Expertenkarte. Oben links: Karte der ungestemmten Texte (schwarz) mit der Expertenkarte (rot) prokrustet; oben rechts: Karte der gestemmten Texte (schwarz) mit der Expertenkarte (rot) prokrustet; unten: ungestemmte Karte (schwarz) mit der gestemmten Karte (rot) prokrustet.

9.4.1 Zusatzvergleich unscharfer/exakter Wortvergleich

Es muss noch geklärt werden, ob sich ein unscharfer Wortvergleich nachteilig auf gestemmte Texte auswirkt. Dazu werden die TargetWords der Stemming-Texte der Variante B einmal nach exakter Übereinstimmung bestimmt und einmal nach unscharfer Wortähnlichkeit. Abbildungen 18 und 19 zeigen die beiden resultierenden Karten: Es sind keine bedeutenden Unterschiede feststellbar, der Stress ist beim unscharfen Wortvergleich etwas höher.

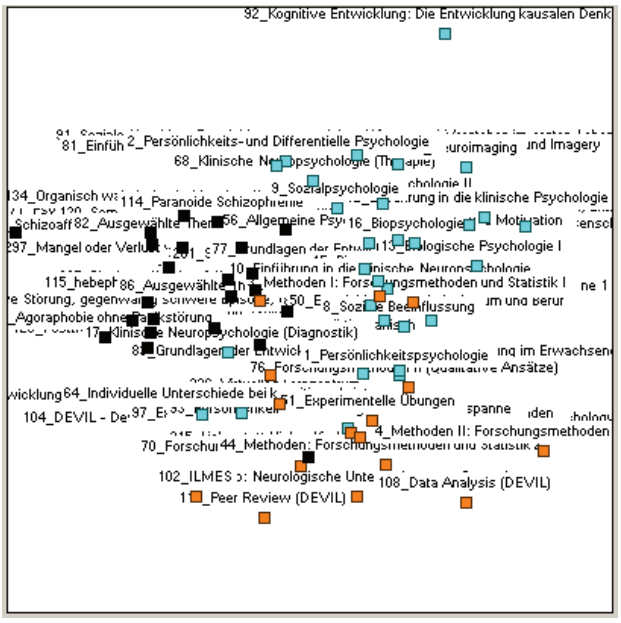


Abbildung 18: exakter Wortvergleich, Stress 0.183

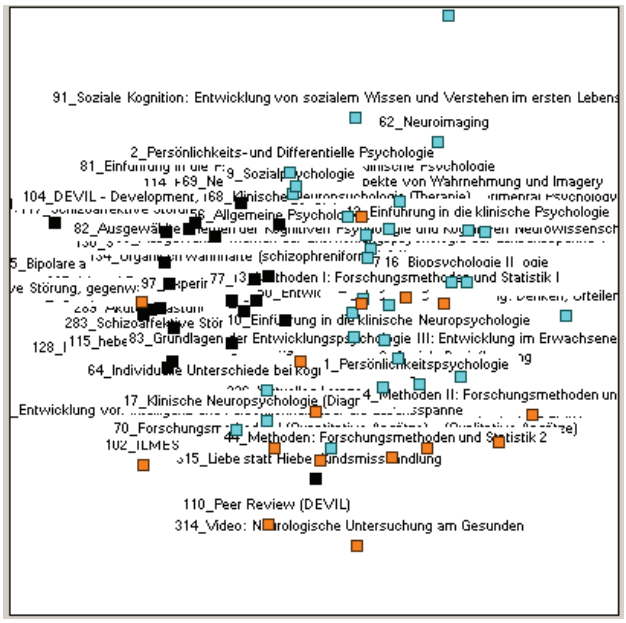


Abbildung 19: unscharfer Wortvergleich, Stress 0.208

Die Prokrustes-Transformation der beiden Karten bestätigt diesen Eindruck (Abb. 20).

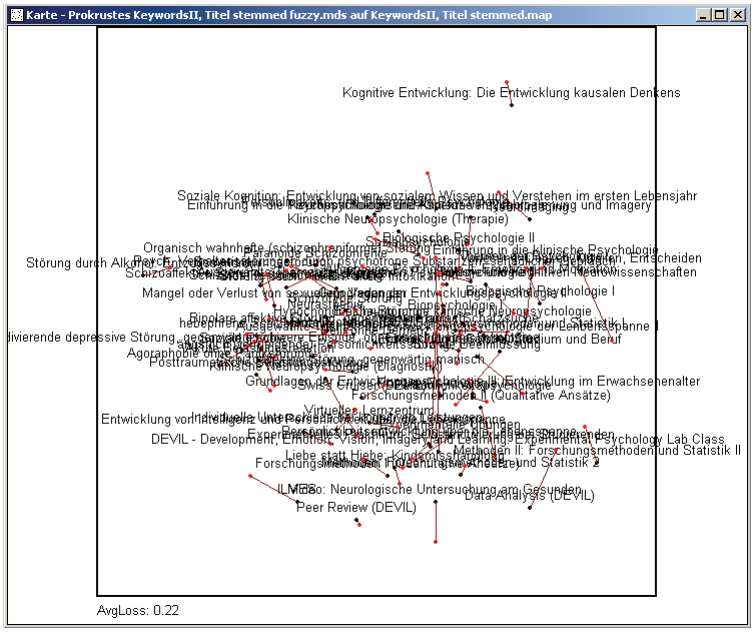


Abbildung 20: Prokrustes-Transformation zwischen exaktem (schwarz) und unscharfem (rot) Wortvergleich

Die Prokrustes-Transformationen mit der Expertenkarte zeigen einen leicht höheren Average Loss für die Karte mit dem unscharfen Wortvergleich.

Somit kann bestätigt werden, dass sich der Aufwand eines unscharfen Wortvergleichs bei gestemmten Texten nicht lohnt. Falls Stemming doch verwendet wird, dann sollten die Keywords mittels exaktem Wortvergleich gesucht werden. Der unscharfe Wortvergleich erhöht nur unnötig das Rauschen in den Daten und kostet zusätzliche Rechenkapazität (Abb. 21).

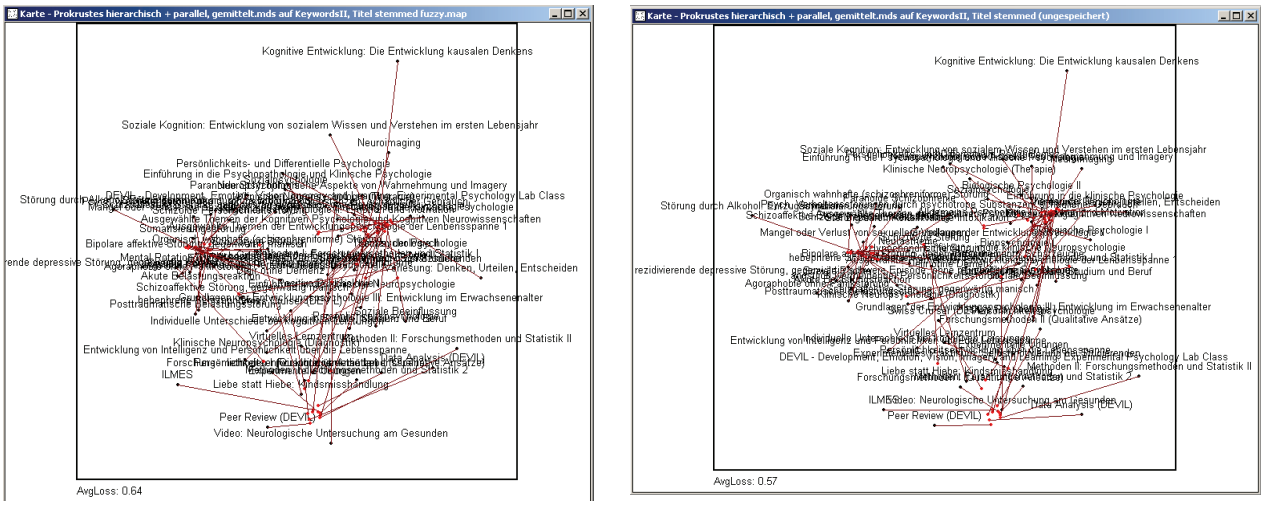


Abbildung 21: Links ist die Prokrustes-Transformation der Expertenkarte mit der HM-Karte (unscharfer Wortvergleich), rechts diejenige mit der HM-Karte (exakter Wortvergleich). Der Average Loss ist beim unscharfen Wortvergleich etwas höher (rot: Items der Expertenkarte; schwarz: Items der HM-Karten).

9.5 Diskussion

Das Stemming hat nur einen marginalen Einfluss auf die Strukturierung der Karten. Der Stress wird geringfügig reduziert. Die Rechen- und Programmierkosten eines Stemmings stehen somit in einem schlechten Verhältnis zum Nutzen und lohnen sich kaum. Wir verzichteten deshalb bei sämtlichen Experimenten in vorliegender Arbeit auf das Stemming.

Teil III Funktionsbeschreibung SemanticMapper

Der SemanticMapper (SM) ist die Softwareumgebung, die wir entwickelt haben, um Experimente und Untersuchungen rund um die Hofmethode durchführen zu können. Das Programm wurde in «Visual Basic .NET» geschrieben (in der Programmierumgebung «Microsoft Visual Studio 2008»[19]). Die Programmierung des SM ist – für eine Einzelperson – ein enorm umfangreiches Projekt. Das Programm besteht in der gegenwärtigen Version 2.84 aus 16'800 Zeilen Code (zum Vergleich: diese Dissertation besteht aus etwa 4'000 Zeilen Text), 457 Funktionen und Eigenschaften. Das GUI verteilt sich auf 8 Fenster, nicht eingerechnet ein Dutzend Dialogfenster.

Der SM ist integraler Bestandteil dieser Dissertation, weshalb seine wichtigsten Funktionen hier beschrieben werden. Das Programm ist aber deutlich «work in progress» – es wird laufend angepasst, umgeschrieben und erweitert. Der vorliegende Teil «Funktionsbeschreibung SemanticMapper» ist keine Anleitung, die den User durch das Programm führt, sondern eine Auslegung dessen Funktionen, um einen Eindruck davon zu vermitteln, was im aktuellen Zustand damit möglich ist.

Durch den mehrjährigen Entwicklungsprozess sind Inkonsistenzen in der Benennung verschiedener Bestandteile entstanden. So sind «TargetWords» (Zielwörter) gefundene Stichwörter, während «Keywords» die gesamte Liste aller Stichwörter bezeichnet, allerdings wird das im Code nicht konsequent umgesetzt. Auch deutsch- und englisch-sprachige Bezeichnungen wechseln sich systemlos ab, sowohl im Code, als auch im GUI. Teilweise wird der veraltete Begriff «Textcloud» anstatt «Tagcloud» verwendet.

19 http://www.microsoft.com/visualstudio/

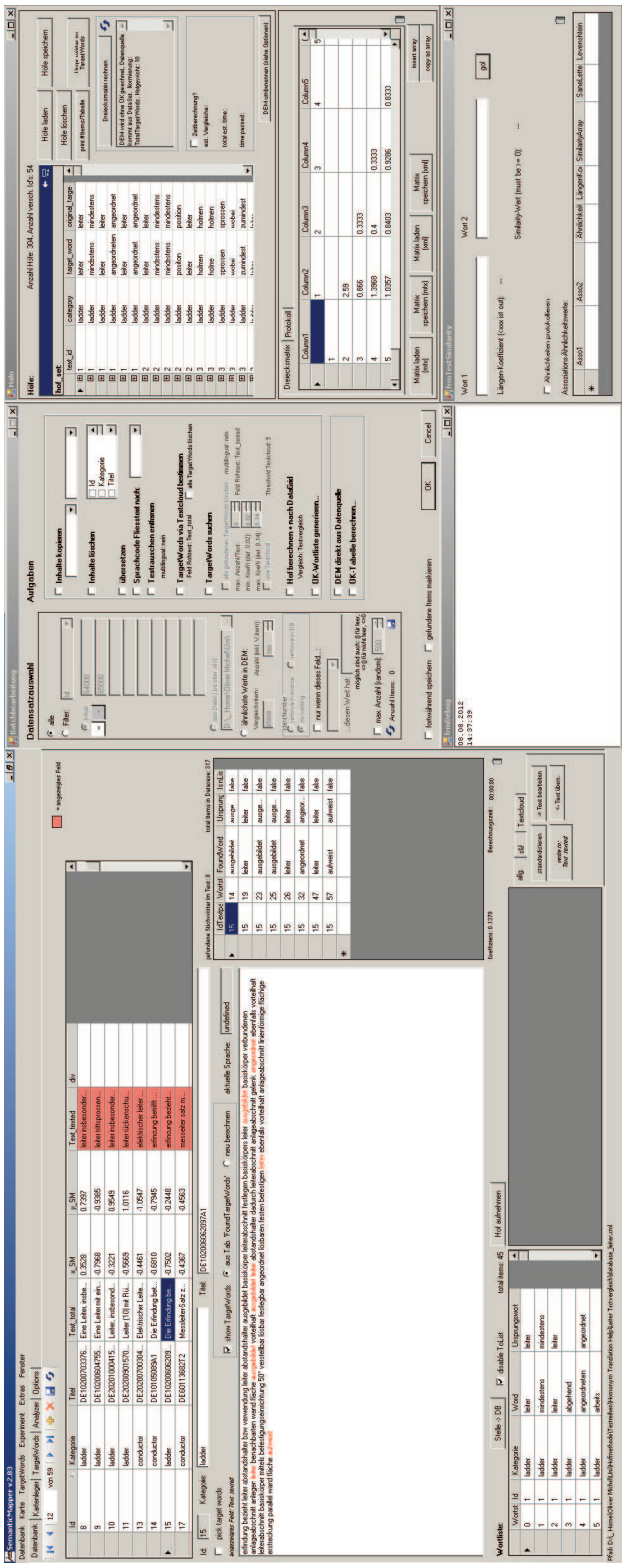


Abbildung 22: Die Oberfläche des SemanticMappers

10 Hauptfenster

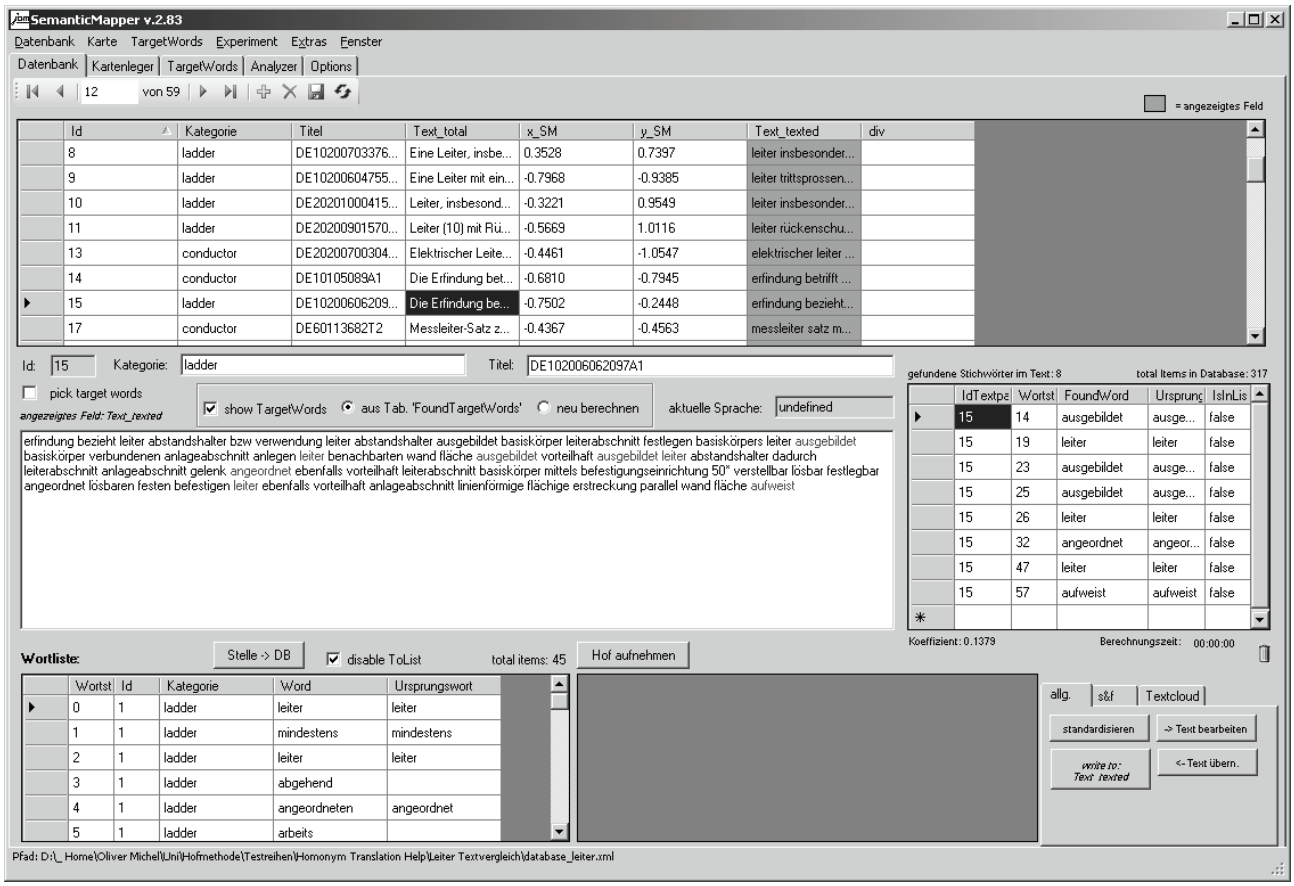


Abbildung 23: Das Hauptfenster – in der Abbildung ist aktuell das Register «Datenbank» angewählt.

Das Hauptfenster gliedert sich in mehrere Registerblätter, die wiederum verschiedene Bereiche aufweisen. Allen Registerblättern gemeinsam ist eine Menüzeile. Zuerst wird diese Menüzeile besprochen, anschliessend die Registerblätter.

10.1 Die Menüzeile

10.1.1 Menü «Datenbank»

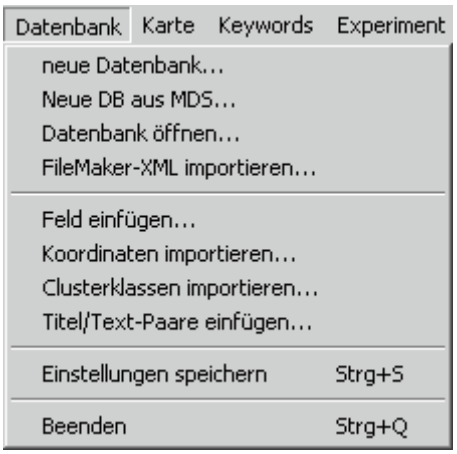


Abbildung 24: Menü «Datenbank»

Menüpunkt «neue Datenbank...»

Es wird eine neue, leere Datenbank erstellt, in der nur die unbedingt nötigen Felder vorhanden sind. Der Speicherort der DB muss in einem folgenden Dialogfeld eingegeben werden. Die DB wird im XML-Format gespeichert.

Menüpunkt «neue DB aus MDS...»

Erstellt eine neue Datenbank auf der Basis einer MDS-Datei, die aus dem ProDax stammt. Die MDS-Datei enthält die Namen und Koordinaten der Items einer Karte.

Menüpunkt «Datenbank öffnen...»

Öffnet eine SM-Datenbank (XML-Dateiformat).

Menüpunkt «FileMaker-XML importieren...»

Konvertiert eine FileMaker-Datenbank, die zuvor aus dem FileMaker als XML exportiert wurde, in eine SM-Struktur.

Menüpunkt «Feld einfügen...»

Fügt ein neues Datenfeld in die aktuelle Datenbank ein.

Menüpunkt «Koordinaten importieren...»

Liest die Koordinaten aus einer MDS-Datei und schreibt sie in die Spalten (die in den Optionen angegeben wurde) der betreffenden Items.

Menüpunkt «Clusterklassen importieren...»

Schreibt die Daten einer Hierarchischen Clusteranalyse (HCA) in die Spalte (welche in den Optionen angegeben wurde) der korrespondierenden Items. Die Daten der HCA müssen sich in der Zwischenablage

befinden (beispielsweise aus Excel kopiert), wobei die Id-Nummern der Items in der ersten Spalte sind und die Clusternummern in der zweiten.

Menüpunkt «Titel/Text-Paare einfügen...»

Dient dem schnellen Erstellen einer neuen Datenbank: Sich in der Zwischenablage befindlichen Titel-Text-Paare werden der Datenbank hinzugefügt. Die Paare können z. B. aus Excel kopiert werden. Mit «Text» ist der Fliesstext/Rohtext des Items gemeint.

Menüpunkt «Einstellungen speichern»

Speichert sämtliche Einstellungen des GUI im Benutzerverzeichnis des Betriebssystems.

Menüpunkt «Beenden»

Speichert die Datenbank und die Einstellungen, dann wird das Programm beendet.

10.1.2 Menü «Karte»

Dieses Menü bezieht sich auf die Karte, welche im Register «Kartenleger» des Hauptfenster angezeigt wird. Diese Einträge sind selbsterklärend.

Abbildung 25: Menü «Karte»

10.1.3 Menü «Keywords»

Menüpunkte «Keywords importieren/exportieren...»

Liest die Keywords aus oder schreibt sie in eine .targetwords-Datei (XML-Format) (s. Kap. 10.4, Hauptfenster, Register «Keywords»).

Abbildung 26: Menü «Karte»

Menüpunkt «Keywords übersetzen...»

Bestimmt die Sprache der aktuellen Keywords und übersetzt sie in die vier komplementären Sprachen (DE, EN, FR, IT und SP). Im aktuellen Arbeitsverzeichnis (s. Optionen) werden entsprechende .targetwords-Dateien erstellt.

Menüpunkte «AllWords importieren/exportieren...»

Mit «AllWords» ist die Wortliste gemeint (s. Kap. 10.4, Hauptfenster, Register «Keywords»), die sämtliche Wörter des angegebenen Datenbestandes umfasst.

Menüpunkt «AllWords nach Keywords»

Kopiert die AllWords-Liste nach Keywords.

10.1.4 Menü «Experiment»

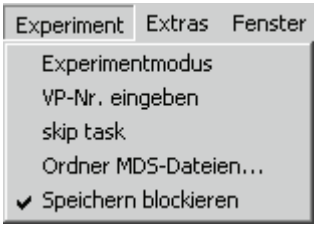


Abbildung 27: Menü «Experiment»

Mit dem SM können Experimente mit VPn durchgeführt werden. Die Experimente sind in der Form, dass der VP eine Frage präsentiert wird, die sie unter Zuhilfenahme einer Karte beantworten soll (s. Kap. 10.3.3, Unterregister «Experiment»). Dazu soll sie auf ein Item in der Karte klicken, worauf der zugrundeliegende Text angezeigt wird. Falls es sich nicht um den gesuchten Lösungstext handelt, muss sie weiterklicken. Die benötigten Klicks werden gezählt und in der verbundenen Access-Datenbank gespeichert. Diese Access-DB ist ein Relikt aus den Anfängen des SM und wird nur noch für die sogenannte «Versuchsanordnung» benötigt, einer Tabelle, in der die VP-Nummern, die Fragen- und Kartennummern, sowie die Anzahl Klicks gespeichert sind. Die Kartennummer bezieht sich auf den Namen der MDS-Datei, welche im aktuellen Durchgang angezeigt wird. Die Fragennummer bezieht sich auf die Auswahl einer hartcodierten (!) Frage. Die Datenbank kann direkt über das Programm Microsoft Access bearbeitet werden oder über das Fenster «Versuchsanordnung» (s. Kap. 13.3, Fenster «Versuchsanordnung»). Der Pfad zur Access-Datenbank ist ebenfalls hartcodiert.

Menüpunkt «Experimentmodus»

Schaltet den SM in den Experiment-Modus: Das Unterregister «Experiment» im Register «Kartenleger» wird angewählt und verschiedene Steuerelemente werden ausgeblendet.

Menüpunkt «VP-Nr. eingeben»

Eingabe der Nummer der aktuellen VP, damit die entsprechenden Daten aus der DB gelesen werden können.

Menüpunkt «skip task»

Aktuelle Aufgabe wird übersprungen (wird nur für Testzwecke benötigt).

Menüpunkt «Ordner MDS-Dateien...»

Hier kann der Ordner angegeben werden, in dem sich die verschiedenen Koordinatendateien befinden.

Menüpunkt «Speichern blockieren»

Das Speichern von Daten (in die Tabelle «Versuchsanordnung») wird unterbunden.

10.1.5 Menü «Extras»

Menüpunkt «Batchbearbeitung...»

Öffnet das Fenster «Batchbearbeitung» (s. Kap. 12, Fenster «Batchbearbeitung»).

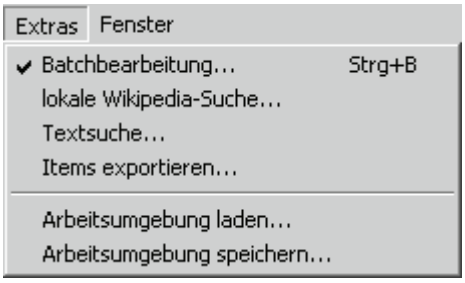


Abbildung 28: Menü «Extras»

Menüpunkt «lokale Wikipedia-Suche...»

Öffnet das Fenster «Wikipedia-Suche» (s. Kap. 13.5, Fenster «Suchmaske Wikipedia»).

Menüpunkt «Textsuche...»

Öffnet das Fenster «Text-Suche» (s. Kap. 13.6, Fenster «Textsuche»).

Menüpunkt «Items exportieren...»

Öffnet eine Dialogbox, in der die Felder markiert werden können, deren Inhalt in eine CSV-Datei geschrieben werden soll.

Menüpunkte «Arbeitsumgebung laden/speichern...»

Zwei praktische Menüpunkte (die von Anfang an hätten programmiert sein müssen!): Die gesamte Arbeitsumgebung (Einstellungen, Fensterpositionen, Dateipfade, Keywords, Hofsets, Inhalte von

Suchmasken...) lässt sich damit speichern und wiederherstellen. Somit ist ein Wechsel zwischen verschiedenen Versuchsreihen rasch möglich.

10.1.6 Menü «Fenster»

Die verschiedenen Fenster des SM lassen sich mit den entsprechenden Menüeinträgen ein- und ausblenden.

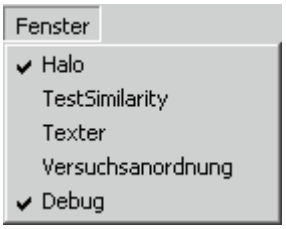


Abbildung 29: Menü «Extras»

10.2 Hauptfenster, Register «Datenbank»

Im oberen Drittel wird die aktuelle Datenbank angezeigt. Durch Doppelklick auf einen Spaltenkopf erscheint der Inhalt dieses Feldes (des markierten Datensatzes) im mittleren Bereich («angezeigtes Feld»). Normalerweise ist das der entrauschte (im SM-Vokabular «getextete/texted») Text.

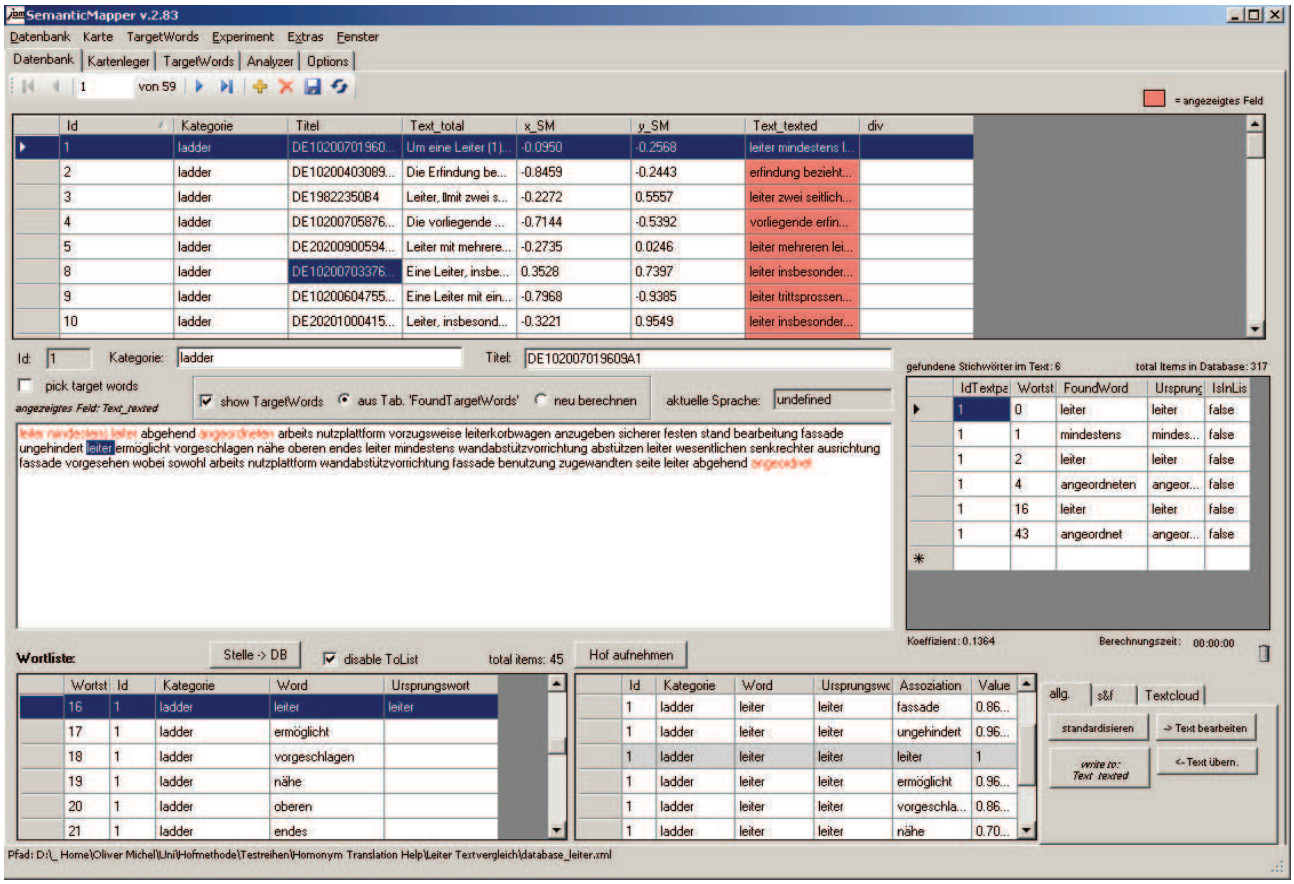


Abbildung 30: Hauptfenster, Register «Datenbank»

Falls die TargetWords schon bestimmt wurden, werden diese rot markiert. Zusätzlich werden sie in der Tabelle Mitte rechts aufgelistet. Die Zellen der Datenbank können markiert und mit Rechtsklick in die Zwischenablage kopiert werden. Auf diesem Weg können auch aus Excel kopierte Werte eingefügt werden. Sämtliche Daten der Datenbank können direkt in der Maske bearbeitet werden.

Links unten befindet sich die Wortliste: Sämtliche Wörter des aktuellen Textes sind hier aufgeführt, ergänzt um eine Id-Nummer, der Wortstelle innerhalb des Textes, der Kategorie (sofern vorhanden) und um das originale Keyword, falls es sich bei diesem Wort um ein TargetWord handelt. Wird ein Wort im angezeigten Feld doppelgeklickt, wird es in der Wortliste markiert und der Hof erstellt. Dieser erscheint rechts von der Wortliste.

Ist die Checkbox «pick target words» markiert, wird mit einem Doppelklick auf ein Wort dieses in die Keyword-Liste aufgenommen.

10.2.1 Unterregister «allg.»

Rechts unten befindet sich das Unterregister «allg.» mit folgenden Buttons:

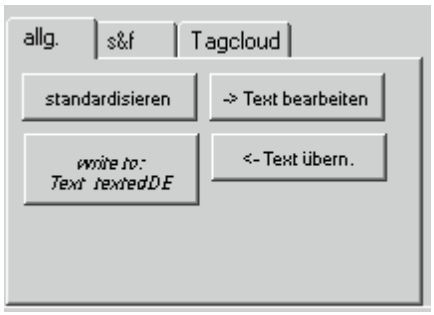


Abbildung 31: Unterregister «allg.»

Button «standardisieren»

Falls im angezeigten Feld der unbearbeitete Fliesstext gezeigt wird, kann dieser hier entrauscht werden.

Button «write to: [angegebenes Feld]»

Der angezeigte Text wird in das Feld geschrieben, welches in den Optionen unter «Feld mit entrauschtem Text» angegeben wurde.

Button «Text bearbeiten»

Öffnet den angezeigten Text im Fenster «Texter» (s. Kap. 13.2, Fenster «Texter»), um ihn zu bearbeiten.

Button «Text übern.»

Übernimmt den fertig bearbeiteten Text aus dem Fenster «Texter» in das «angezeigte Feld».

10.2.2 Unterregister «s&f»

Hier kann im angezeigten Feld nach einem Wort oder Teilwort gesucht werden. Fundorte werden markiert, auf Wunsch alle gleichzeitig.

Abbildung 32: Unterregister «s&f»

10.2.3 Unterregister «Tagcloud»

Hier wird das API der Tagcloud-Schnittstelle (s. Kap. 16, Tagcloud-Verfahren von Semager) angesprochen. Der Text, welcher sich im «angezeigten Feld» befindet, wird an das API geschickt, welches die «wichtigsten» Wörter, zusammen mit einem Wichtigkeitswert, retourniert. Die Maximalanzahl, sowie der Schwellwert können hier angegeben werden.

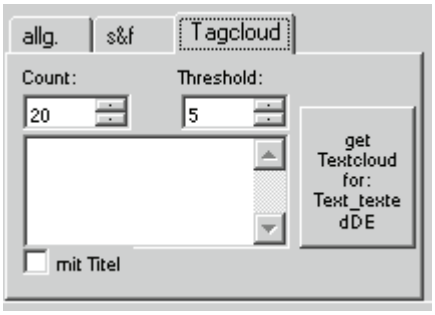


Abbildung 33: Unterregister «Tagcloud»

10.3 Hauptfenster, Register «Kartenleger»

In diesem Register finden alle Operationen im Zusammenhang mit den Karten statt. Wird ein Item in der Karte angeklickt, erscheint dessen Text in der rechten Fensterhälfte. Die Operationsgruppen sind wieder in Register unterteilt.

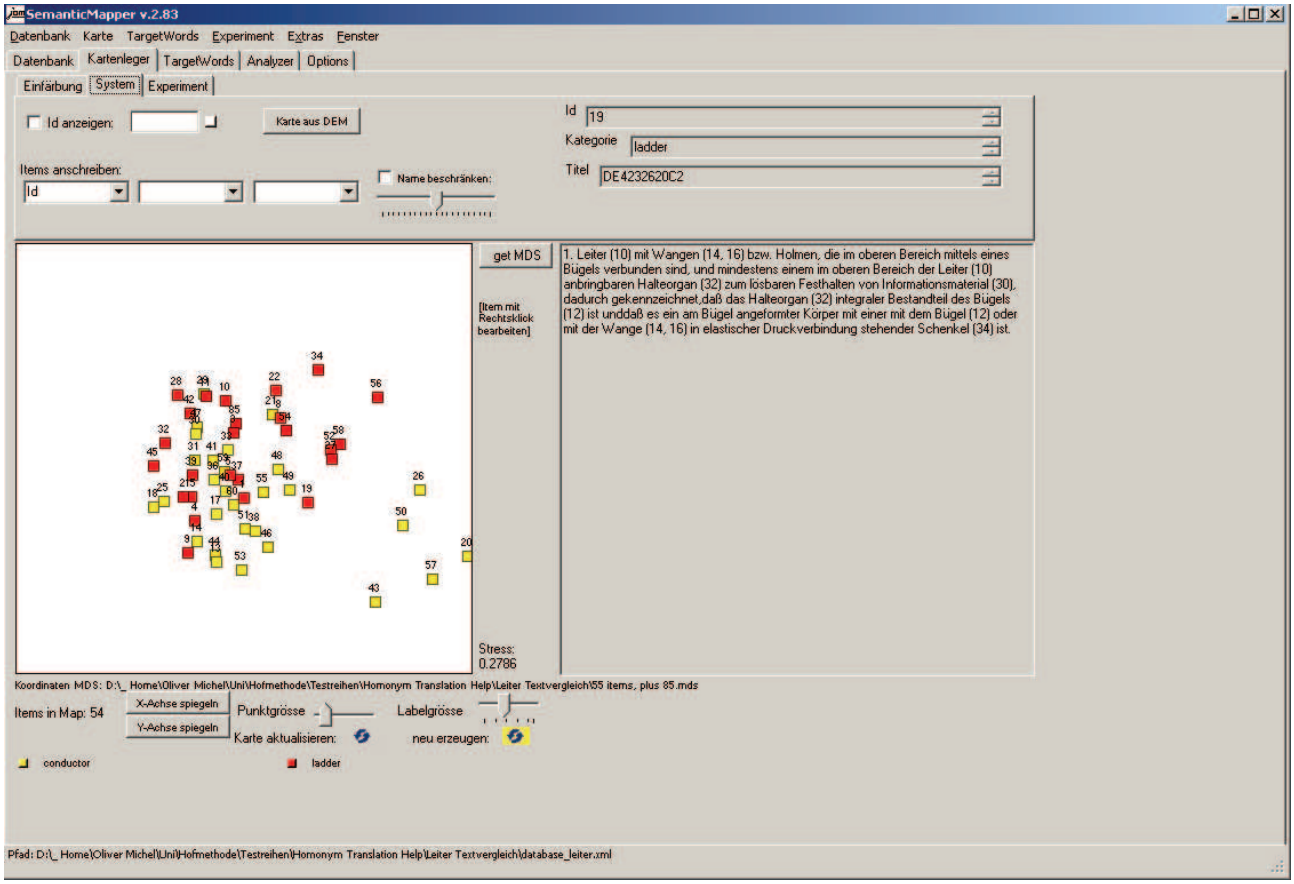


Abbildung 34: Hauptfenster, Register «Kartenleger» (mit ausgewähltem Unterregister «System»)

10.3.1 Unterregister «Einfärbung»

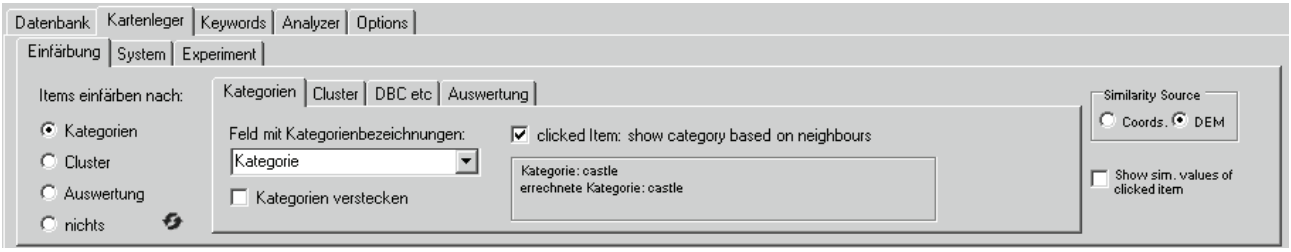


Abbildung 35: Register «Kartenleger», Unterregister «Einfärbung» (mit angewähltem Unterunterregister «Kategorien»)

Links kann die Art der Einfärbung ausgewählt werden. Beim Anklicken wird das entsprechende Unterunterregister rechts davon aktiviert.

Ganz rechts aussen kann die Quelle der Ähnlichkeitswerte angegeben werden (Koordinaten oder Dreiecksmatrix).

Ist die Checkbox «Show sim. values of clicked item» aktiviert, wird nach einem Klick auf ein Item dessen Ähnlichkeitswerte zu den anderen Items durch rote Striche angezeigt, deren Intensität mit den Werten korreliert (Abb. 36).

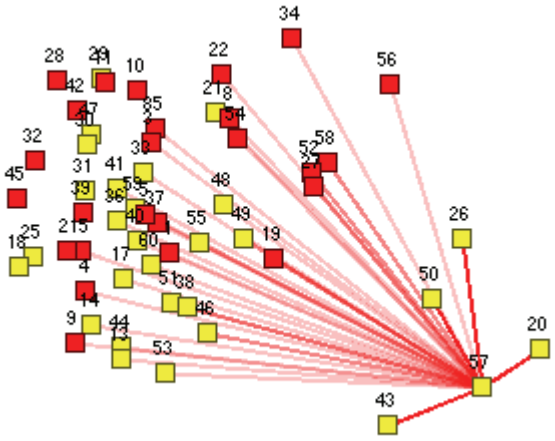


Abbildung 36: Karte, wenn die Checkbox «Show sim. values of clicked item» aktiviert ist.

Unterregister «Einfärbung»: Unterunterregister «Kategorien»

Das Feld mit der einzufärbenden Kategorie muss im Dropdown ausgewählt werden. Unterhalb der Karte erscheint dann eine Legende der vorkommenden Kategorien (s. Abb. 34). Durch einen Klick auf den farbigen Button hinter der Kategorie kann der Farbwert verändert werden.

Ist die Checkbox «clicked Item: show category based on neighbours» aktiviert, wird im Feld darunter die zugewiesene, sowie die errechnete Kategorie eines Homonyms gezeigt (s. Kap. 26, Homonymdisambiguierung: Anwendungsbeispiel).

Unterregister «Einfärbung»: Unterunterregister «Cluster»

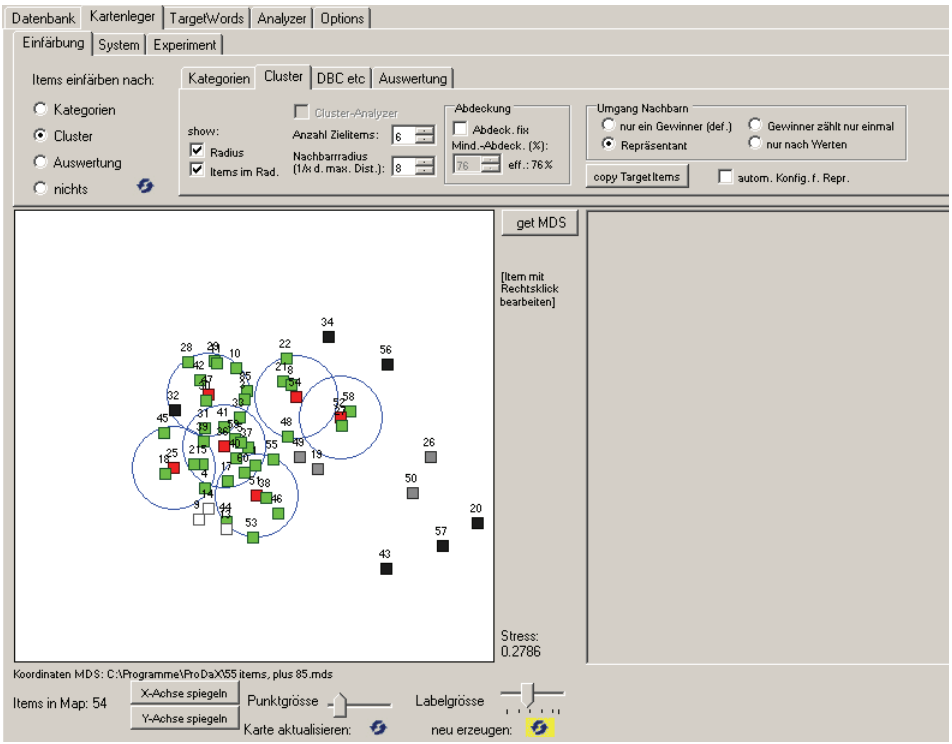


Abbildung 37: Unterregister «Einfärbung»: Unterunterregister «Cluster». In der Karte sind sechs Clustermittelpunkte rot markiert. Grün markiert sind die Items, die sich innerhalb des Sichtbarkeitshorizontes befinden (dieser ist blau eingezeichnet). Die Grauabstufung der restlichen Items repräsentiert deren «Nachbarkeitswert».

Hier werden sämtliche Operationen rund um die Clusteritemidentifikation gesteuert. Die Anzahl der benötigten Zielitems, wie auch die Grösse des Nachbarradius/Sichtbarkeitshorizontes (s. Kap. 17, Repräsentanten-Algorithmus) können bestimmt werden, ebenso, wie die durch die Clusteritems benötigte Abdeckung. Verschiedene Berechnungsvarianten für die Nachbarbestimmung stehen rechts zur Auswahl (s. Anhang A3, Zielitems in der NMDS nach Verteilungen bestimmen).

Die Checkbox «Cluster Analyzer» bestimmt die Anzahl Clusteritems und den Sichtbarkeitshorizont automatisch, ebenso wie dies die Checkbox «autom. Konfig. f. Repr.» für die Repräsentanten macht.

Der Button «copy TargetItems» kopiert die Id-Nummern der Clusteritems in die Zwischenablage.

Unterregister «Einfärbung»: Unterunterregister «DBC etc.»

Auf dieses Register wird hier nicht eingegangen, da das «Distribution Based Coloring» (Ryf, 2008) in vorliegender Arbeit nicht behandelt wird.

Unterregister «Einfärbung»: Unterunterregister «Auswertung»

Wird ein Experiment durchgeführt, können hier beliebige Fragen-Karten-Kombinationen eingestellt werden, die dann im Register «Experiment» erscheinen.

10.3.2 Unterregister «System»

Hier können die Labels der Items verändert werden. Zudem kann ein bestimmtes Item farblich hervorgehoben werden («Id anzeigen»).

Durch Klick auf den Button «Karte aus DEM» wird der Inhalt der Dreiecksmatrix (s. Kap. 11, Fenster «Halo») zum ProDax geschickt, welches die NMDS rechnet und abspeichert. Sobald das geschehen ist, werden die Koordinaten in die Datenbank des SM importiert und die Karte angezeigt.

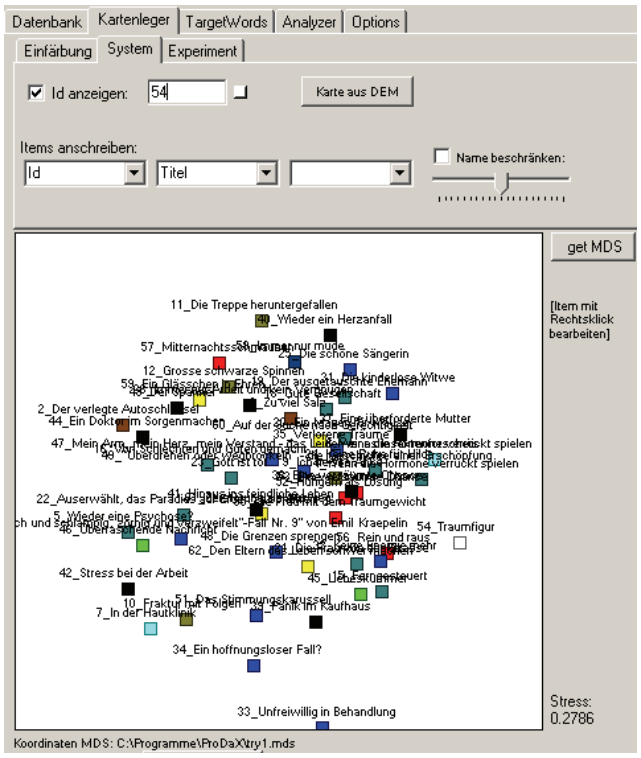


Abbildung 38: Unterregister «System»: Hier können die Items angeschrieben werden.

10.3.3 Unterregister «Experiment»

Dieses Register dient der Anzeige der Fragen, die bei einem Experiment mit VPn benötigt werden (s. Kap. 24, Wikipedia-Experiment).

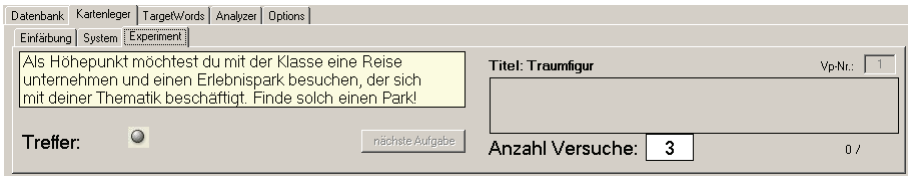


Abbildung 39: Fragendisplay bei einem Experimentsdurchgang

10.4 Hauptfenster, Register «Keywords»

Dieser Register dient der Kompilierung von Keywordlisten. Es stehen zwei Methoden zur Auswahl: KeywordsII (s. Kap. 15, KeywordII-Analyse) und Wortfrequenzmethode (s. Kap. 14, Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient). Die Parameter auf der linken Seite dienen alle der KWII-Methode.

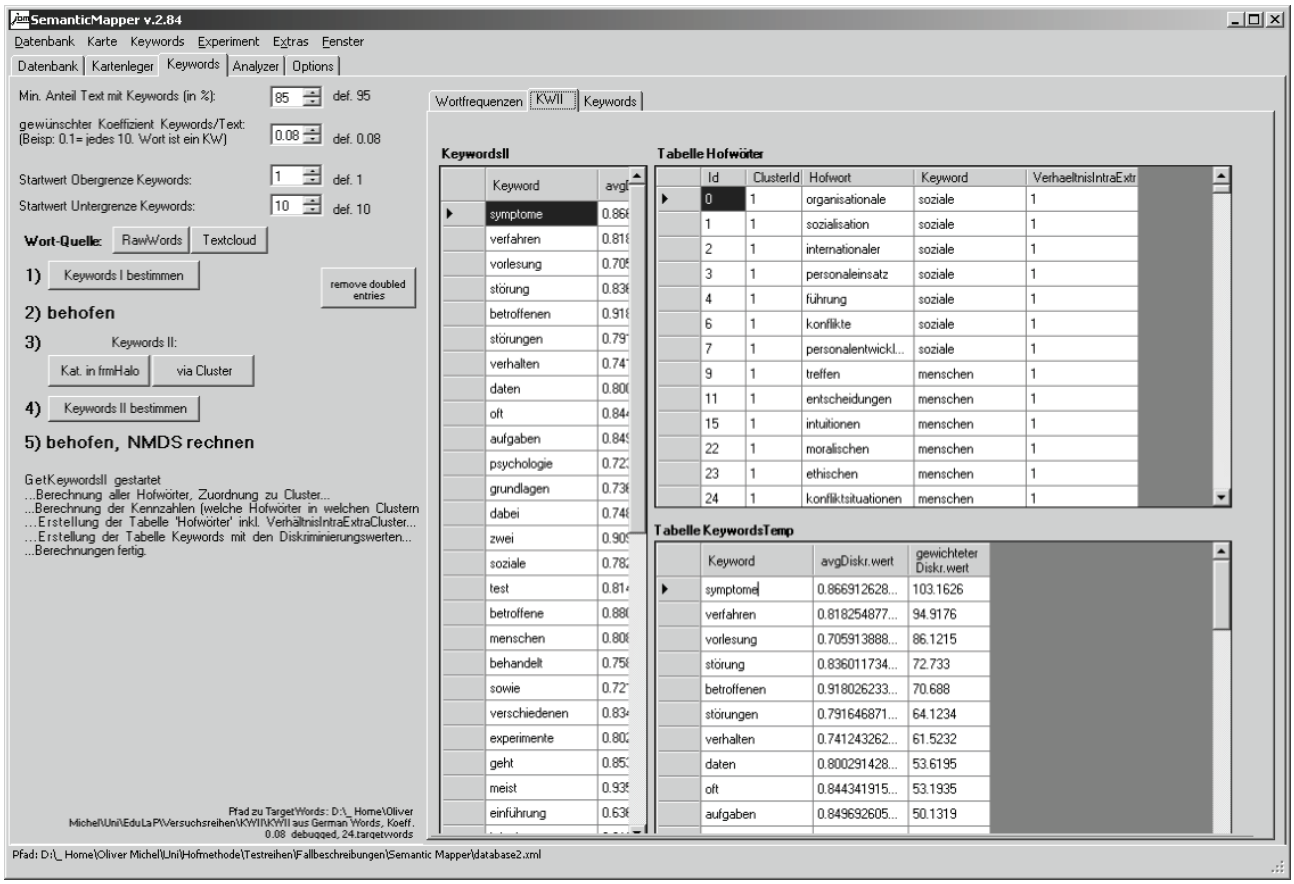


Abbildung 40: Hauptfenster, Register «Keywords» (mit ausgewähltem Unterregister «KWII»)

Im Unterregister «Keywords» wird die Keywordliste und die umfassende Liste aller Wörter der Datenbank dargestellt. Im Unterregister «KWII» werden Zwischenergebnisse aus dem Berechnungsprozess angezeigt.

Das Unterregister «Wortfrequenzen» dient der zweiten Methode der Keywordbestimmung. Im oberen Teil wird der Berechnungsprozess durch den Button «Get Keywords from HCA» gestartet. Voraussetzung ist, dass die Clusternummern in dem in den Optionen angegebenen Datenfeld vorhanden sind. Die Wortfrequenzen werden ausgegeben und können mit dem Button «zu Keywords» in die Keywordliste übertragen werden.

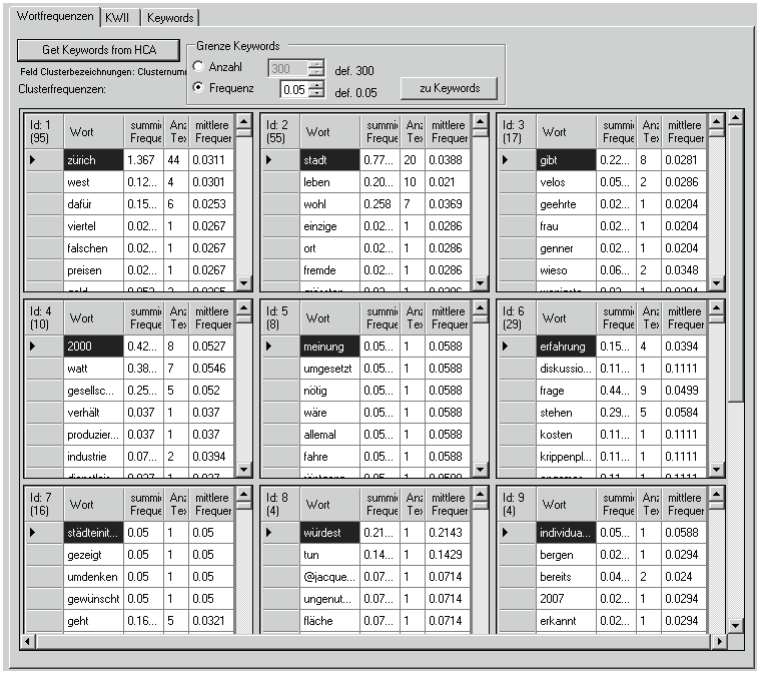


Abbildung 41: Hauptfenster, Register «Keywords» (mit ausgewähltem Unterregister «Wortfrequenzen»)

10.5 Hauptfenster, Register «Analyzer»

Dieses Register dient dem einfacheren Analysieren der Höfe. Leider ist die Programmierung nicht sehr weit fortgeschritten. In der oberen Zeile können die Id-Nummern der zu untersuchenden Texte eingegeben werden. Sind die entsprechenden Checkboxen aktiviert, werden die TargetWords rot und die gemeinsamen TargetWords blau markiert. Mit Drag'n'Drop können die zu untersuchenden TargetWords aus dem linken Bereich in den rechten gezogen werden, um deren Hofwörter zu analysieren.

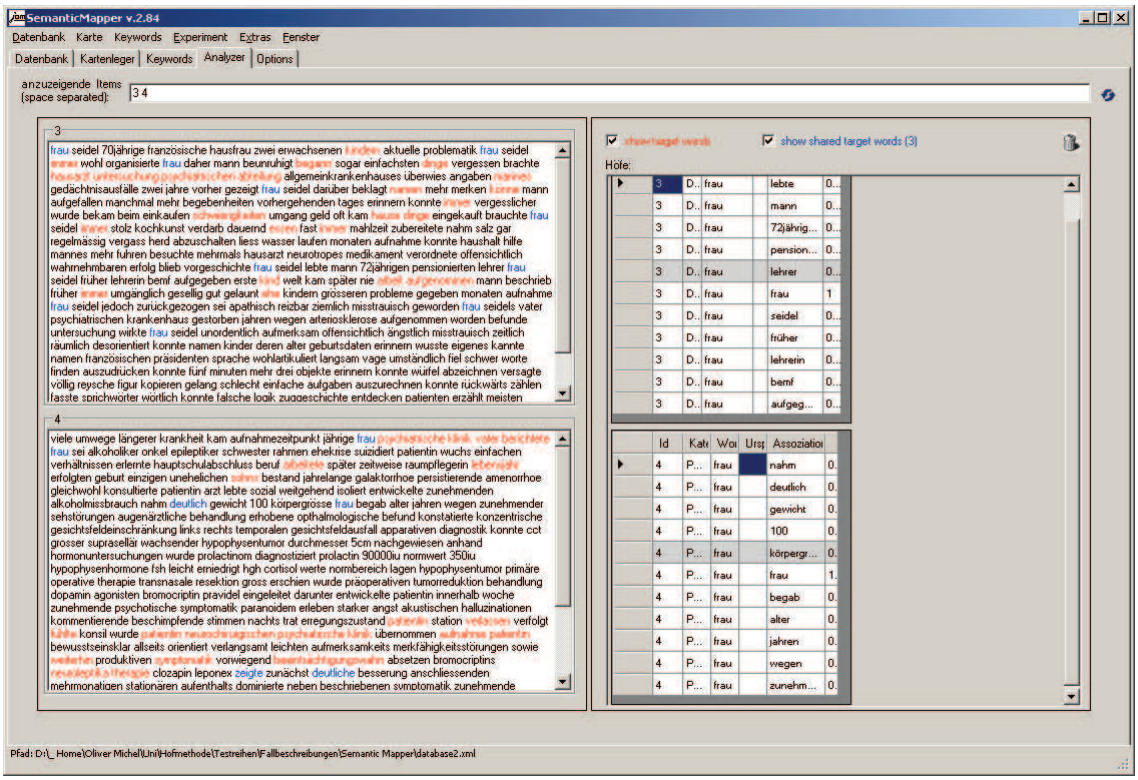


Abbildung 42: Hauptfenster, Register «Analyzer», mit zwei exemplarisch herausgezogenen Höfen

10.6 Hauptfenster, Register «Options»

Im Register «Options» werden an zentraler Stelle die Optionen des SM zusammengefasst. Die meisten Einstellungen sind selbsterklärend. Hier werden nur diejenigen aufgeführt, deren Bedeutung nicht aus der Bezeichnung erschlossen werden kann.

Bereich «Hofmethode»: «threshold similarity»

Schwellwert für die Levenshtein-Funktion (s. Kap. 2.1.5, Unscharfe Ähnlichkeit), die die Ähnlichkeit der Assoziationswörter (die Wörter im Hof) vergleicht.

Bereich «Hofmethode»: «fuzzy Wortvergleich»

Gibt an, ob ein unscharfer Keywordvergleich durchgeführt werden soll, das heisst ob die TargetWords exakt einem Keyword entsprechen müssen oder ob eine gewisse Ähnlichkeit (siehe oben) genügt.

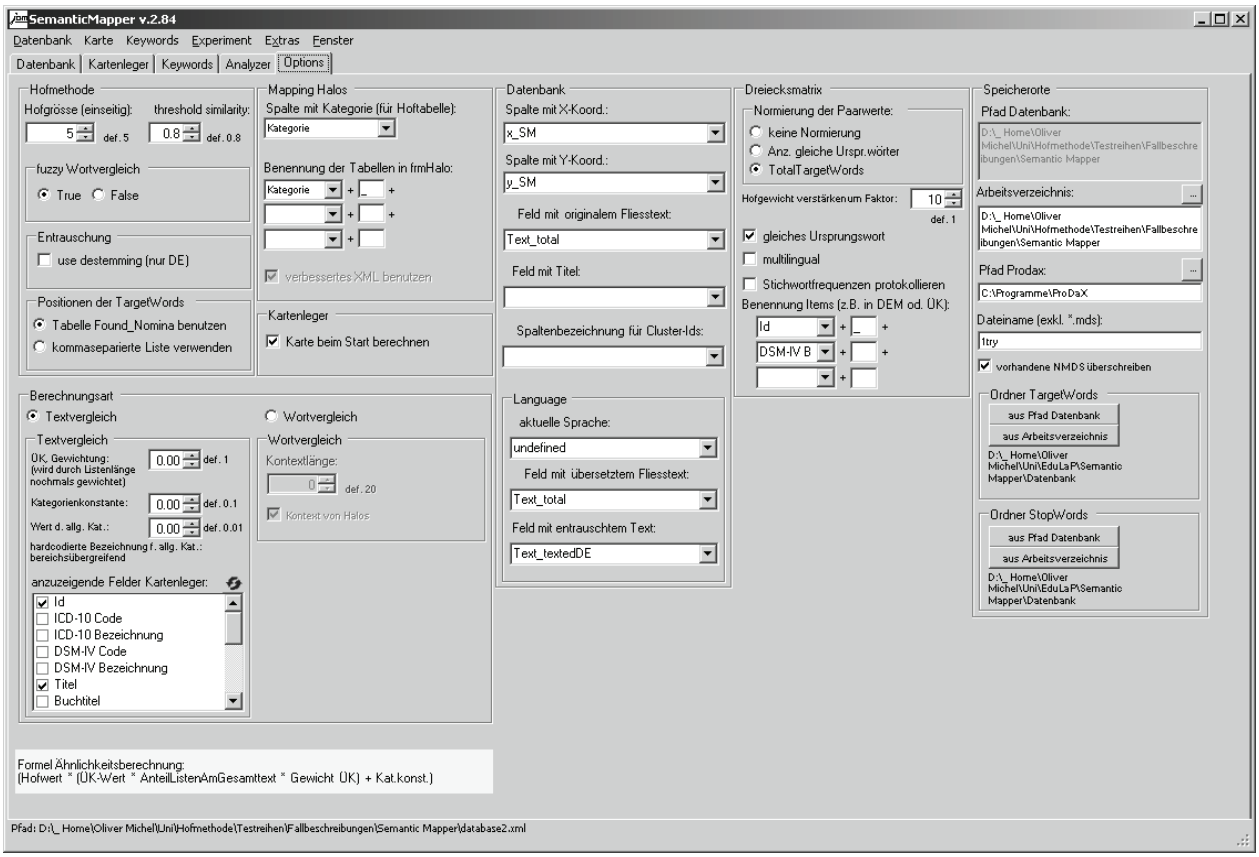


Abbildung 43: Hauptfenster, Register «Optionen»

Bereich «Berechnungsart»

Hier kann der Berechnungsmodus bestimmt werden, der angibt, ob Texte miteinander verglichen werden oder einzelne Wörter. In vorliegender Arbeit wird nur mit dem Textvergleich gearbeitet. Beim Wortvergleich würde die Verwendung eines bestimmten Wortes innerhalb eines einzigen Textes verglichen werden.

Bereich «Language»

Diese Angaben stehen im Zusammenhang mit der Multilinguality (s. Kap. 21, Multilinguality). Falls nicht mehrsprachig gearbeitet wird, muss im Feld «aktuelle Sprache» der Eintrag «undefined» ausgewählt werden.

Bereich «Dreiecksmatrix»

Die Normierung und das Hofgewicht sind in den Kapiteln 5 (Normierung der Textähnlichkeitswerte: SharedTargetWords vs. TotalTargetWords) und 6 (Hofgewichtung) beschrieben.

Checkbox «gleiches Ursprungswort»: Ist diese Option aktiviert, findet der Hofvergleich nur statt, wenn beide TargetWords dasselbe ursprüngliche Keyword haben (diese Option hebelt somit den «fuzzy Wortvergleich» vor Berechnung der DEM wieder aus). Ist die Option nicht aktiviert, wird für jedes TargetWord-Paar dessen Ähnlichkeit berechnet und beim Unterschreiten eines «threshold similarity»-Wertes der Vergleich verworfen.

Checkbox «multilingual»: Falls aktiviert, wird für jede Sprache (DE, EN, FR, IT, SP) eine separate DEM gerechnet und dann gemittelt (s. Kap. 21, Multilinguality).

Checkbox «Stichwortfrequenzen protokollieren»: Die Frequenzen der Stichwörter können zu Auswertungszwecken in die Tabelle «Protokoll» des Fensters «Halo» geschrieben. werden Dieses Feature wird in vorliegender Arbeit nicht genutzt.

11 Fenster «Halo»

Dieses Fenster stellt einen Kontrollpunkt im Berechnungsprozess der Ähnlichkeitsvergleiche dar – es wäre nicht unbedingt nötig, da die Karten direkt aus der Datenquelle heraus berechnet und erstellt werden können (s. Kap. 12, Fenster «Batchbearbeitung»). Trotzdem ist es für den Entwicklungsprozess praktisch, denn so lassen sich beispielsweise die berechneten Höfe anschauen und speichern. Aufwendige Berechnungen lassen sich damit später wiederverwerten.

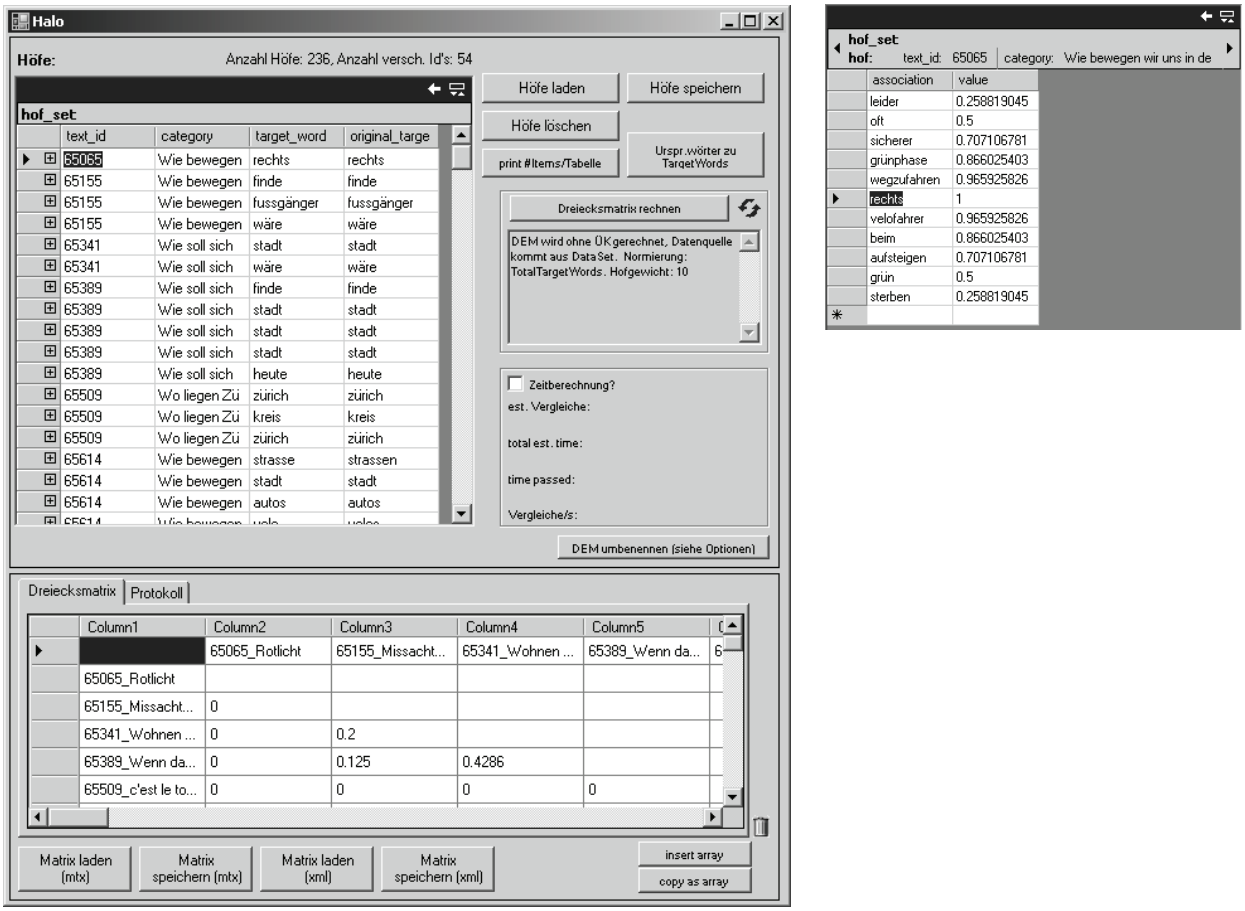


Abbildung 44: Links: Das Fenster «Halo» mit einem kompletten Hofset und der berechneten DEM unten; rechts: Ein einzelner Hof des Sets ist aufgeklappt.

Im linken Bereich ist das sogenannte «Hofset» dargestellt: Ein «Hofset» beinhaltet sämtliche Höfe der zu vergleichenden Texte. Die Höfe gelangen entweder manuell über das Hauptfenster oder automatisiert über die Batchbearbeitung dorthin. Die Höfe können als .halos-Dateien (XML-Format) gespeichert und gelesen werden.

Der Button «Dreiecksmatrix rechnen» startet die Berechnung der Textähnlichkeiten gemäss den gewählten Optionen. Zur Orientierung sind die wichtigsten Optionen im Feld darunter nochmals zusammengefasst.

Ist die Checkbox «Zeitberechnung» aktiviert, wird vor der eigentlichen Berechnung eine Simulation gestartet, um den voraussichtlichen Zeitbedarf abschätzen zu können.

Die DEM im unteren Bereich lässt sich als .mtx- oder als .xml-Datei speichern (und laden). Das MTX-Format wird von ProDax verwendet. Die vom SM erstellten MTX-Dateien können also direkt im ProDax verarbeitet werden. Das XML-Format ist nötig, falls die Itemanzahl gross sind (ca. >1'000 Items). Das verwendete DataView-Element der VisualBasic-Umgebung kann diese Datenmenge nicht fassen. Werden grosse ÜK-Tabellen benötigt (wie in Kap. 28, Semantische Strukturierung eines Diskussionsraums), werden diese deshalb als XML-Datei gespeichert und also solche in den Speicher geladen, aber nicht im DataView-Element angezeigt.

Ebenfalls im unteren Bereich befindet sich die Stichwortfrequenzprotokollierung, welche in vorliegender Arbeit aber nicht eingesetzt wird.

12 Fenster «Batchbearbeitung»

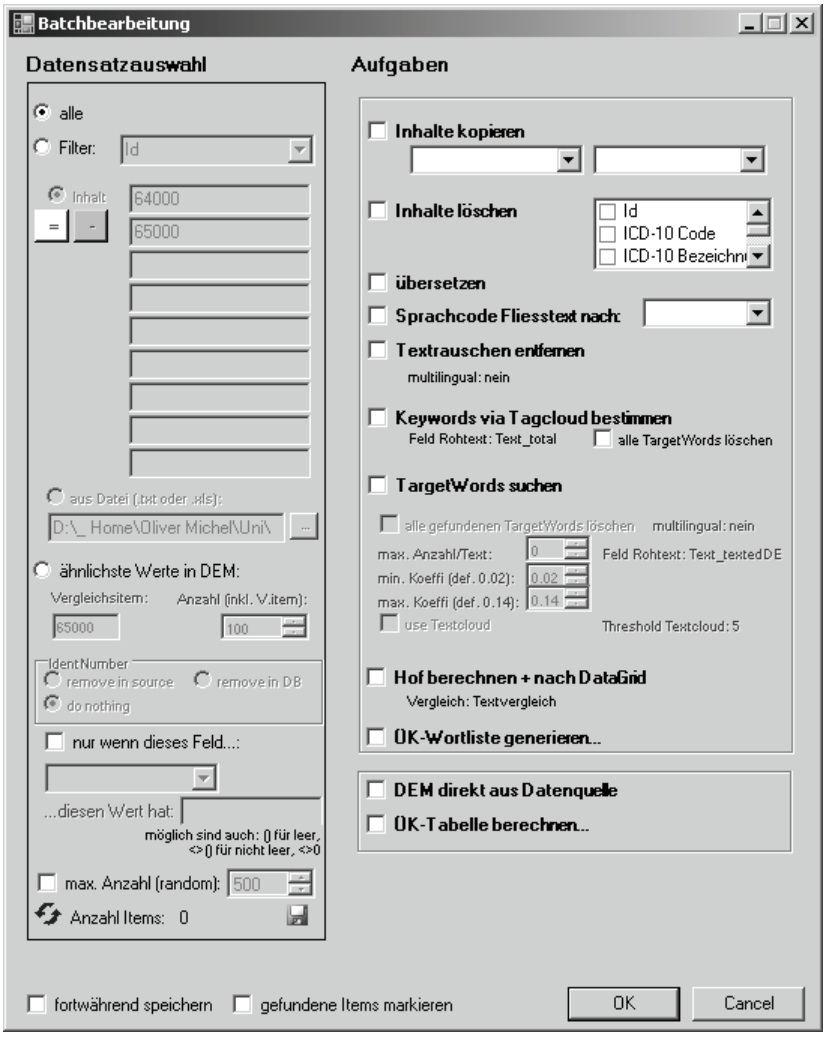


Abbildung 45: Das Fenster «Batchbearbeitung»

Mit Hilfe dieses Fensters lassen sich bestimmte Aufgaben auf ausgewählte Datensätze anwenden. Meistens macht man sämtliche Berechnungen über dieses Fenster.

In der linken Hälfte können die Datensätze ausgewählt werden, in der rechten Hälfte werden die auszuführenden Aufgaben ausgewählt.

12.1 Fenster «Batchbearbeitung», Bereich «Datensatzauswahl»

Option «alle»

Sämtliche Datensätze der Datenbank werden für die Bearbeitung ausgewählt. Trotzdem kann die Auswahl durch die beiden Checkboxen «nur wenn dieses Feld...» und «max. Anzahl» (siehe unten) wieder eingeschränkt werden.

Option «Filter»

Für das ausgewählte Datenfeld können bis zu neun Vergleichswerte eingegeben werden oder es wird ein Bereich (z. B. Id-Nummern 200 – 400) angegeben.

Option «aus Datei (.txt oder .xls)»

Hier kann eine Liste mit den Id-Nummern aus einer einfachen Text- oder Exceldatei importiert werden.

Option «ähnlichste Werte in DEM»

Aus der aktuell geöffneten DEM (im Fenster «Halo») werden die Items mit den höchsten Paarwerten zum angegebenen Item herausgesucht. Die DEM darf dabei sowohl im MTX-Format als auch im XML-Format vorhanden sein.

Checkbox «nur wenn diese Feld...»

Hier kann die bis anhin getätigte Auswahl wieder eingeschränkt werden.

Option «max. Anzahl (random)»

Möchte man eine Maximalanzahl von ausgewählten Items nicht überschreiten, kann hier eine Limite gesetzt werden. Die Items werden per Zufall aus dem ausgewählten Pool ausgesucht.

Ein Klick auf das Diskettensymbol speichert die Id-Nummern der aktuellen Auswahl in einer anzugebenden Textdatei.

12.2 Fenster «Batchbearbeitung», Bereich «Aufgaben»

Checkbox «übersetzen»

Die Sprache des Rohtextes (das korrespondierende Feld muss in den Optionen angegeben werden) wird automatisiert festgestellt, anschliessend wird der Text in die vier übrigen der fünf Sprachen DE, EN, FR, IT und SP übersetzt.

Checkbox «Textrauschen entfernen»

Der Fliesstext wird mit den Wörtern der Stoppwortliste (s. Kap. 13.2, Fenster «Texter») gefiltert und entrauscht. Falls in den Optionen «multilingual» aktiviert ist, geschieht die Entrauschung für alle fünf Sprachen, wobei sich die Stoppwortlisten im Arbeitsverzeichnis befinden müssen.

Checkbox «Keywords via Tagcloud bestimmen»

Die Rohtexte werden an das Tagcloud-API (s. Kap. 16, Tagcloud-Verfahren von Semager) gesendet, welches die «wichtigsten Wörter» gemäss Einstellungen (s. Kap. 10.2, Hauptfenster, Register «Datenbank») retourniert.

Checkbox «TargetWords suchen»

Die entrauschten Texte werden mit der Keywordliste verglichen und die TargetWords bestimmt. Es gelten die Einstellungen, die in den Optionen gemacht wurden. Zusätzlich können hier diverse Parameter eingegeben werden, was die Anzahl und Dichte der TargetWords pro Text betrifft.

Checkbox «Hof berechnen + nach DataGrid»

Je nach dem in den Optionen gewählten Vergleichsmodus werden die Höfe in unterschiedlicher Weise berechnet und in das «Hofset» im Fenster «Halo» (s. Kap. 11, Fenster «Halo») übertragen. Beim Wortvergleich wird zusätzlich zum Hof der benachbarte Textausschnitt des TargetWords gespeichert. Dieser Ausschnitt wird beim Klick auf das Item in der Karte in der rechten Hälfte angezeigt.

Checkbox «ÜK-Wortliste generieren...»

Berechnet eine ProDax-kompatible Wortliste aller ausgewählten Items, die als Textdatei an einem auszuwählenden Ort gespeichert wird. Diese Wortliste kann in ProDax als «Quelle für Assoziationsdaten» eingelesen werden.

Checkbox «DEM direkt aus Datenquelle»

Erstellt die Dreiecksmatrix direkt – ohne Umweg über das «Hofset» – und zeigt sie im Fenster «Halo» an.

Checkbox «ÜK-Tabelle berechnen...»

Berechnet den Überlappungskoeffizienten aller ausgewählten Items und speichert die MTX-Datei am anzugebenden Ort.

13 Übrige Fenster

13.1 Fenster «TestSimilarity»

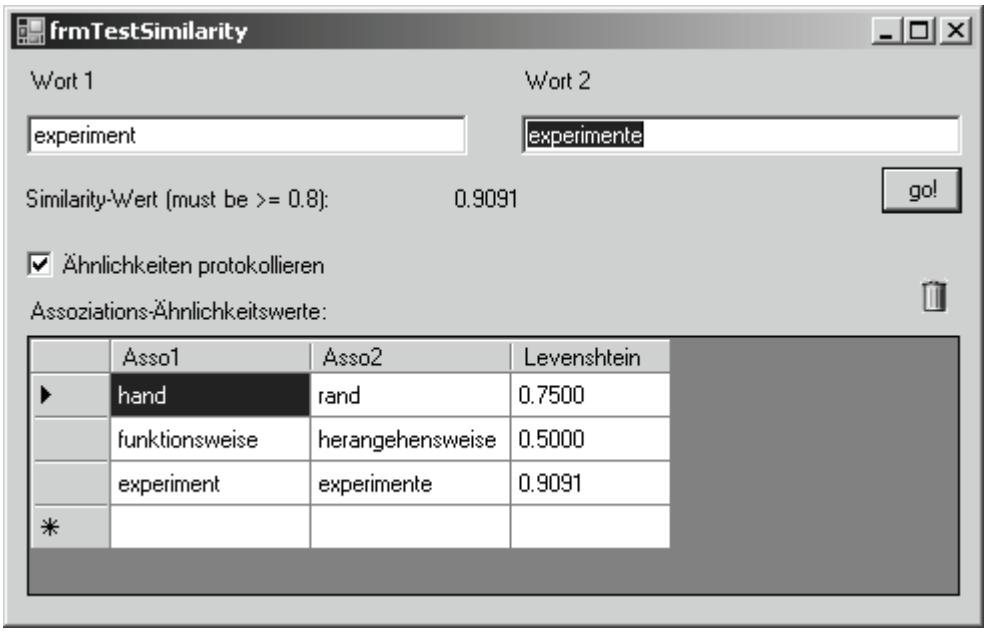


Abbildung 46: Das Fenster «TestSimilarity»: Hier lassen sich manuelle Ähnlichkeitsvergleiche von Wortpaaren durchführen.

In diesem Fenster können Wortpaare manuell auf ihre Ähnlichkeit (nach Levenshtein) überprüft werden. Falls die Checkbox «Ähnlichkeiten protokollieren» aktiviert ist, werden nicht nur die manuell eingegebenen Wortpaare protokolliert, sondern sämtliche Hofpaare, die auch während eines Hofvergleichs anfallen, der aus dem Fenster «Batchbearbeitung» gestartet wurde.

13.2 Fenster «Texter»

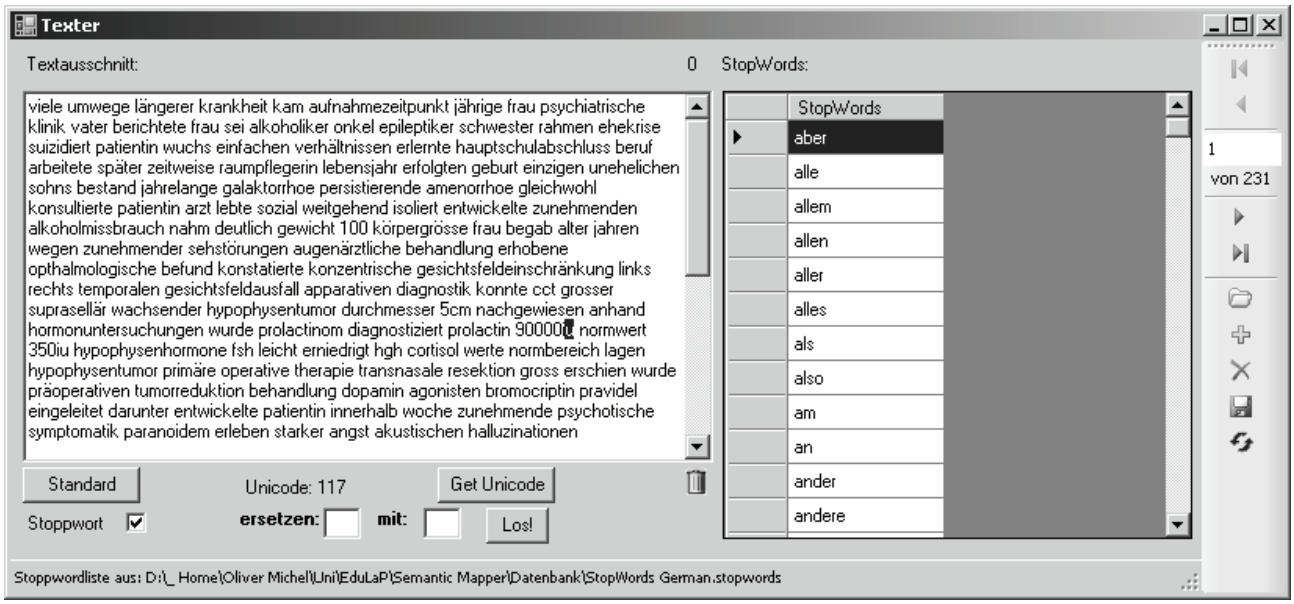


Abbildung 47: Das Fenster «Texter» wird meistens für die Bearbeitung der Stoppwortliste genutzt.

Im Fenster «Texter» können einerseits Textbearbeitungen durchgeführt werden, andererseits wird hier die Stoppwortliste zusammengestellt. Über die Iconliste rechts aussen können .stopwords-Dateien (im XML-Format) oder Textdateien importiert und gespeichert werden.

13.3 Fenster «Versuchsanordnung»

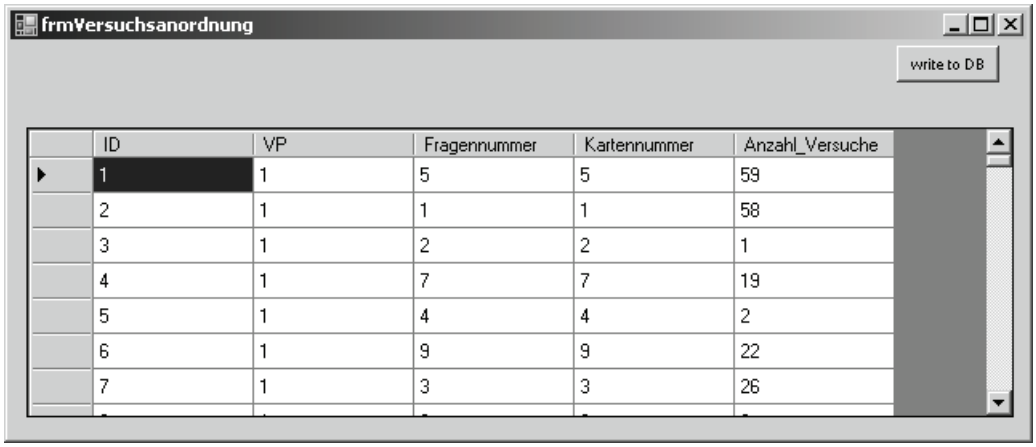


Abbildung 48: Das Fenster «Versuchsanordnung» beinhaltet die Konfiguration für die Experimentsdurchgänge.

In diesem Fenster können die Daten für die Experimentsdurchführung (s. Kap. 10.1.4, Menü «Experiment») bearbeitet werden. Während einer Experimentsphase kann hier überprüft werden, ob die Anzahl Klicks korrekt erfasst wurde.

13.4 Fenster «Debug»

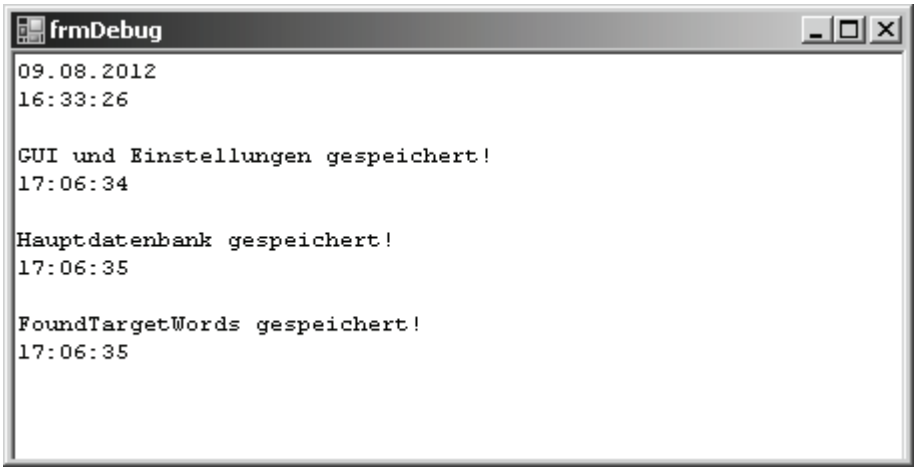


Abbildung 49: Das Fenster «Debug»: Ausgabekonsole für Programmmeldungen

Das Fenster «Debug» dient als Ausgabekonsole für den SM. Informationen und Fehlermeldungen werden hier angezeigt.

13.5 Fenster «Suchmaske Wikipedia»

Dieses Fenster wurde speziell für das Wikipedia-Experiment programmiert (s. Kap. 24, Wikipedia-Experiment). Ein eingegebener Suchbegriff wird im Suchmaschinen-Index gesucht und die korrespondierenden Wikipedia-Artikel werden im unteren Teil angezeigt. Von dort aus können sie in die Datenbank übernommen werden.

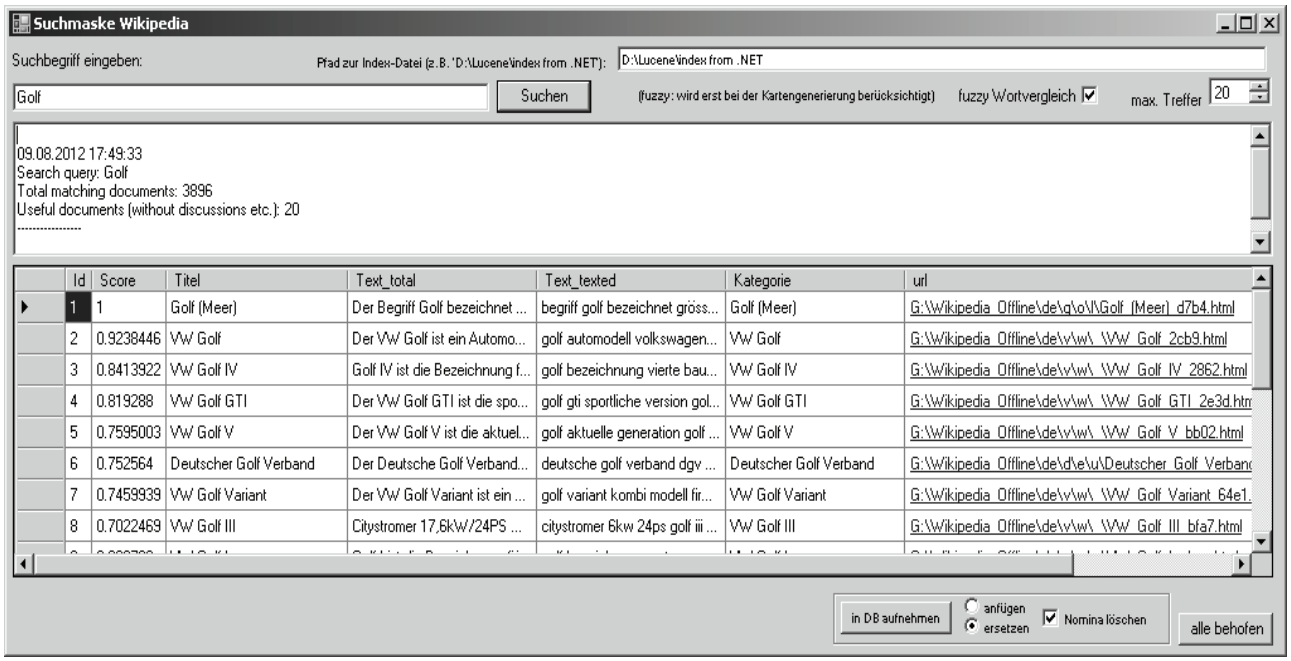


Abbildung 50: Das Fenster «Suchmaske Wikipedia»

13.6 Fenster «Textsuche»

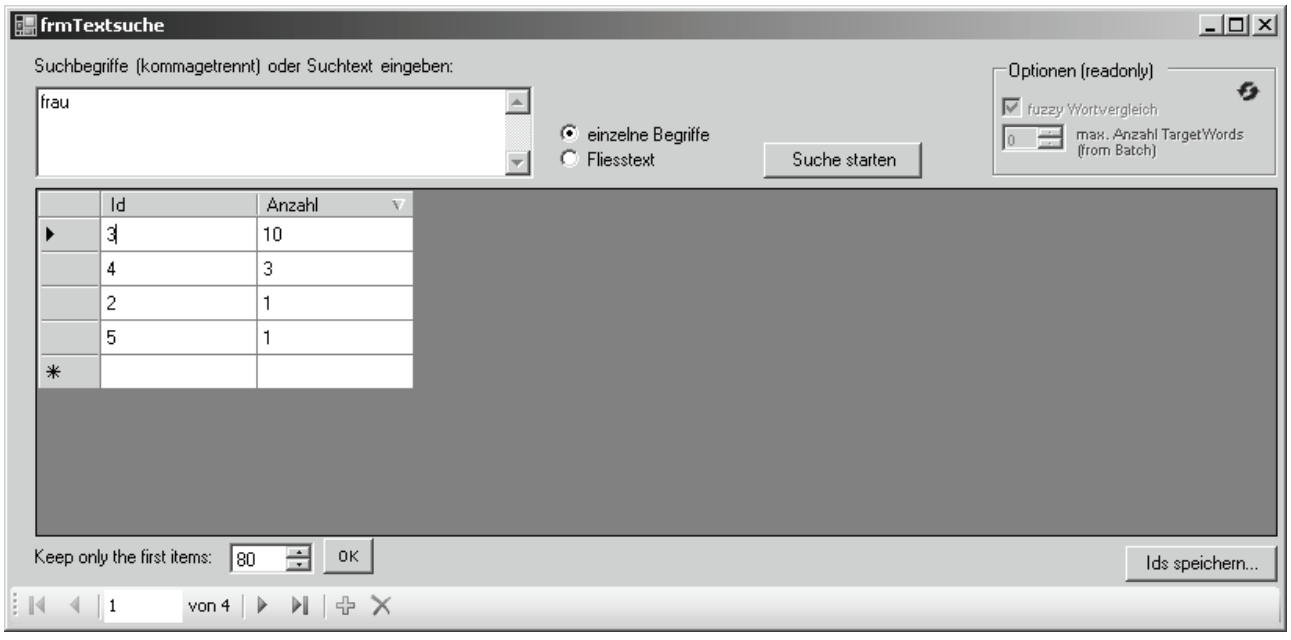
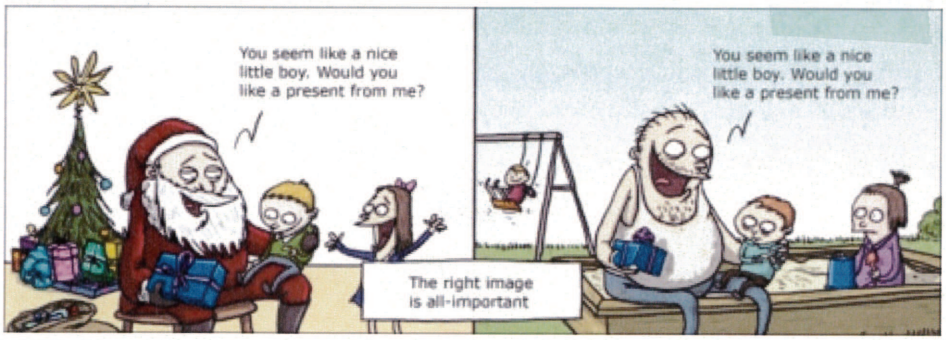


Abbildung 51: Das Fenster «Textsuche»

Mit Hilfe des Fensters «Textsuche» können aus einem grossen Datenbestand Items bestimmt werden, in denen bestimmte Worte vorkommen oder die Ähnlichkeiten zu einem längeren Suchtext haben. Im zweiten Fall wird der eingegebene Suchtext entrauscht und behoft. Das resultierende Hofset wird mit den

(im Fenster «Batchbearbeitung) ausgewählten Items verglichen. Die Id-Nummern und Anzahl, beziehungsweise der Hofwert, werden in die Tabelle im unteren Bereich geschrieben. Die Id-Nummern können in einer Textdatei gespeichert werden, auf die im Fenster «Batchbearbeitung» verwiesen werden kann.

Teil IV Ergänzende Verfahren



14 Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient

14.1 Überblick

Die Wortfrequenzenmethode ist ein von uns entwickelter Algorithmus, der in automatisierter Weise Keywords aus einem gegebenen Textkorpus extrahiert. Bei einer genügend grossen Textbasis funktioniert das Verfahren gut, solange der Umfang der aufzunehmenden Keywords begrenzt wird.

In diesem Kapitel wird der Algorithmus im Detail beschrieben und mit einem alternativen, simplen Verfahren (dem der häufigsten Wörter) verglichen.

14.2 Einleitung

Um die semantischen Relationen von Texten zu berechnen, erstellt die Hofmethode Höfe von bestimmten Keywords. Die Generierung dieser Keywords beschäftigte uns seit Anbeginn der Arbeiten rund um die Hofmethode. Verschiedene Varianten kamen schon zum Einsatz: Bei der Kongress-Karte (s. Kap. 22, Bedeutungsähnlichkeiten von Abstracts: Vier Verarbeitungsebenen) wurden die Keywords aus den Nomina der Beitragstitel manuell ausgewählt; beim edulap-Projekt (s. Kap. 23, Projekt edulap) diente die Regensburger Verbundklassifikation als Datengrundlage; beim Wikipedia-Experiment (s. Kap. 24, Wikipedia-Experiment) waren es die drei Stichwörter «Erde», «Wasser», «Erde/Boden/Grund». Es waren also immer Wörter, die entweder manuell bestimmt werden mussten oder aus einem externen Katalog stammten. Um eine grösstmögliche Flexibilität der Hofmethode zu gewährleisten, ist es aber nötig, die Keywords automatisiert aus dem Textkorpus zu extrahieren.

Ein naheliegendes Verfahren ist, einfach die häufigsten Wörter der Textbasis zu bestimmen und diese als Keywords zu markieren. Wie wir sehen werden, funktioniert das erstaunlich gut.

Das hier diskutierte Verfahren ist aber formal plausibler. Automatisiert werden aus einem Textkorpus diejenigen Stichwörter ausgewählt, welche ein *optimales Separieren von Bedeutungsclustern* ermöglichen. Das Verfahren beruht auf Wortfrequenzen, weshalb es hier Wortfrequenzmethode genannt wird. In diesem Kapitel wird gezeigt, wie sich die Wortfrequenzenmethode gegenüber dem Verfahren der häufigsten Wörter verhält.

14.3 Vorgehen

14.3.1 Konzept der Wortfrequenzmethode

Zuerst müssen die groben Bedeutungscluster identifiziert werden. Dazu bemühen wir den Überlappungskoeffizienten (ÜK): Aus dem Textkorpus wird basierend auf dem ÜK eine Dreiecksmatrix (DEM) gerechnet, daraus liefert eine Clusteranalyse Ähnlichkeitsschwerpunkte. Nun werden die relativen Wortfrequenzen pro Text gerechnet und innerhalb dieser Cluster gemittelt. Anschliessend werden die Abweichungen dieser Frequenzen vom Gesamtdurchschnitt aufsummiert. Dies ergibt hohe Werte für Begriffe, die in einem Cluster oft vorkommen (Werte über dem Durchschnitt), in den anderen aber wenig (Werte unter dem Durchschnitt) und tiefe Werte für Begriffe, die gleichmässig auf alle Texte verteilt sind. Schliesslich wird eine Cut-off-Grenze bestimmt und alle Wörter, die mindestens diesen Wert erreichen, bilden die neue Keywordliste.

14.3.2 Implementierung

Das Verfahren wird anhand eines Datensatzes der Stadtdebatte (s. Anhang, Kap. 2.1, Beschreibung des Webforums Stadtdebatte) getestet. Der gesamte Textkorpus von 2'000 Texten ist allerdings zu umfangreich für eine effiziente Berechnung, deshalb begrenzen wir eine Auswahl. Die Auswahl sollte jedoch so umfassend sein, dass die Diversität der Texte abgebildet werden kann. Bei 2'000 Texten dürfte dies mit einem zufallsbestimmten Subsample von 250 Texten der Fall sein.

Die DEM wird einer agglomerativen, hierarchischen Clusteranalyse unterzogen (Backhaus, 2000). Wir verwenden die Software XlStat[20]. Als Fusionsalgorithmus kommt unweighted-pair group zum Einsatz. Aus den 250 Texten resultieren 13 Cluster. Die Clustergrössen sind 94, 55, 29, 17, 16, 10, 8, 4, 4, 4, 3, 2 und 2 Items. XlStat gibt eine Liste der Text-Ids und der dazugehörigen Clusternummer aus. Diese Liste wird in den SemanticMapper importiert.

20 http://www.xlstat.com

Für jeden Text werden die relativen Wortfrequenzen gerechnet und innerhalb der 13 Cluster gemittelt (s. Abb. 52). Gemäss Konzept werden die Abweichungen dieser Frequenzen vom Gesamtdurchschnitt aufsummiert und in eine Rangfolge gebracht. Als Cut-off-Grenze wählen wir den Frequenzwert 0.1. Dies resultiert in 221 Keywords, welche einer einfachen, manuellen Bearbeitung unterzogen werden (vor allem Präpositionen und Konjunktionen wie «beim» oder «dafür» werden entfernt), worauf 190 Wörter übrig bleiben.

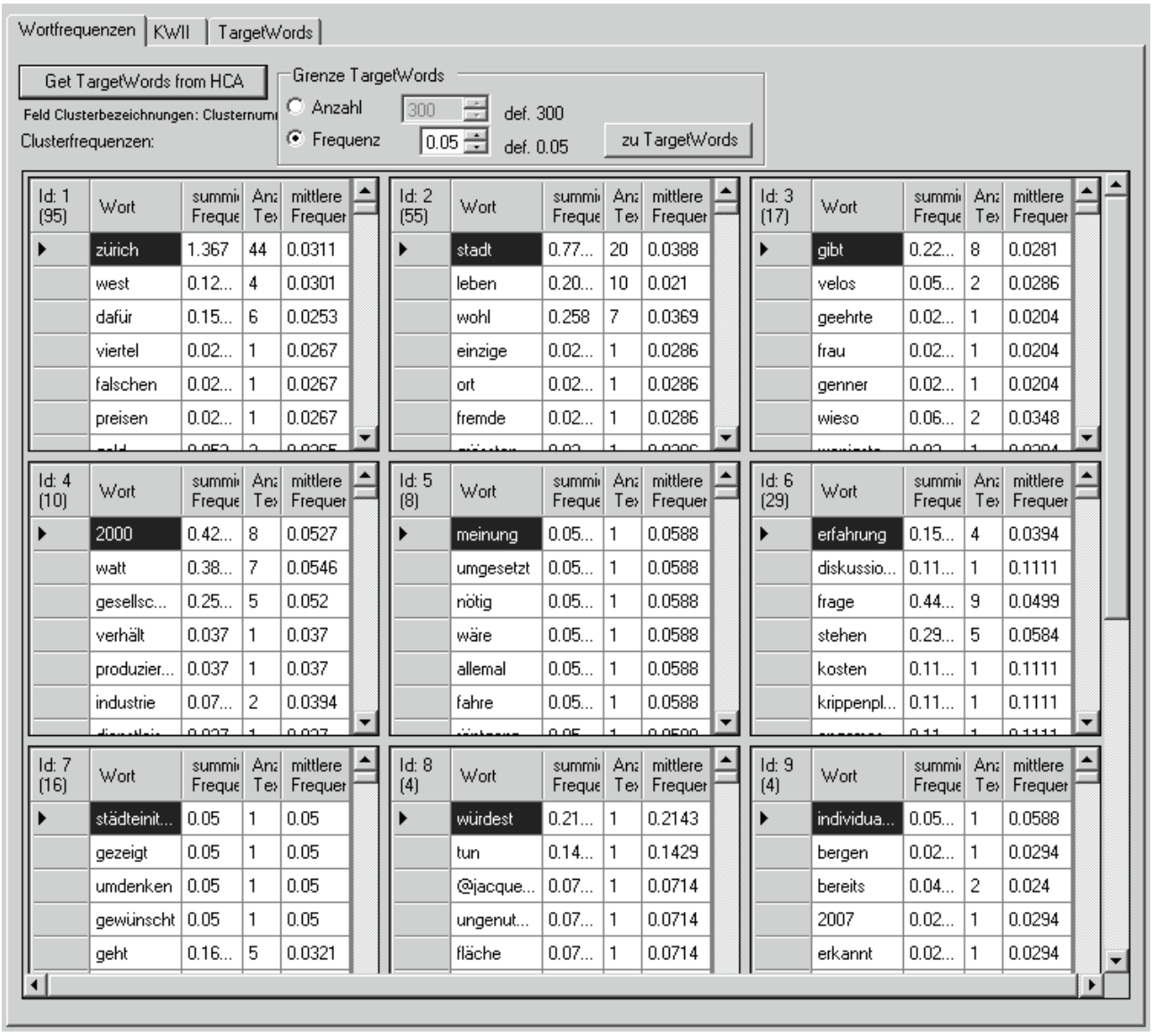


Abbildung 52: Zur Bestimmung der Keywords werden die prozentualen Wortfrequenzen aller Wörter gerechnet, pro Cluster gemittelt und anschliessend die Frequenzabweichung summiert.

14.3.3 Methode

Zunächst wurde ein Testsample von 16 Texten erstellt (s. Anhang 5): Mit den 2'000 Texten des Textkorpus wurde eine DEM basierend auf dem Überlappungskoeffizienten gerechnet. Dann wurde aus drei verschiedenen Foren ein charakteristischer Text ausgewählt und jeweils diejenigen 5 Texte aus der DEM herausgesucht, die den höchsten Paarwert aufwiesen. Zwei Texte wurden danach wieder entfernt: Einer aus formalen Gründen, der andere, weil er inhaltlich sehr verschieden von den restlichen und in diesem Setting nicht von Belang war. Die 16 Texte wurden dann mit unterschiedlichen Keywords behoft:

- mit den 50, resp. 250 häufigsten Wörtern der gesamten Datenbasis
- mit den 50, resp. 190 Wörtern der Wortfrequenzmethode (basierend auf der gesamten Datenbasis)
- mit den 20 häufigsten Wörtern dieser 16 Texte
- mit 20, resp. 50 Wörtern der Wortfrequenzenmethode (aus diesen 16 Texten generiert)

Als Baseline wurde zudem eine Karte berechnet, die auf dem ÜK basiert.

Bei allen Berechnungen wurde keine Kategorienkonstante mit eingerechnet. Der TargetWords-Koeffizient wurde auf den Bereich zwischen 0.02 und 0.14 beschränkt, maximal jedoch 20 TargetWords/Text. Bei Unterschreiten der Limite wurde mittels Tagcloud-Verfahren (s. Kap. 16, Tagcloud-Verfahren von Semager) weitere Keywords gesucht. Durch die Beschränkung des TargetWord-Koeffizienten und die Art, wie die TargetWords innerhalb eines Textes bestimmt werden, kommt eine Zufallskomponente ins Spiel, die auch einen Einfluss auf die Plastizität der Karten hat (s. Unterkap. 14.4.4, Instabilität und Plastizität der Karten).

14.3.4 Visualisierung

Für ein einfacheres Verständnis wird das obige Vorgehen exemplarisch mit den 16 Texten der Stadtdebatte (s. Anhang, Kap. 2.1, Beschreibung des Webforums Stadtdebatte) visualisiert.

Für die 16 Texte wird paarweise der ÜK gerechnet. Skaliert man die resultierende Dreiecksmatrix, sieht die Karte gemäss Abbildung 53 (links oben) aus. Die hierarchische Clusteranalyse erkennt aus der DEM drei Cluster. Diese sind in der Abbildung 53 (rechts oben) eingefärbt (die HCA untersucht die Werte der DEM, nicht die Koordinaten der NMDS, deshalb müssen die Cluster nicht kongruent mit der Karten-darstellung sein).

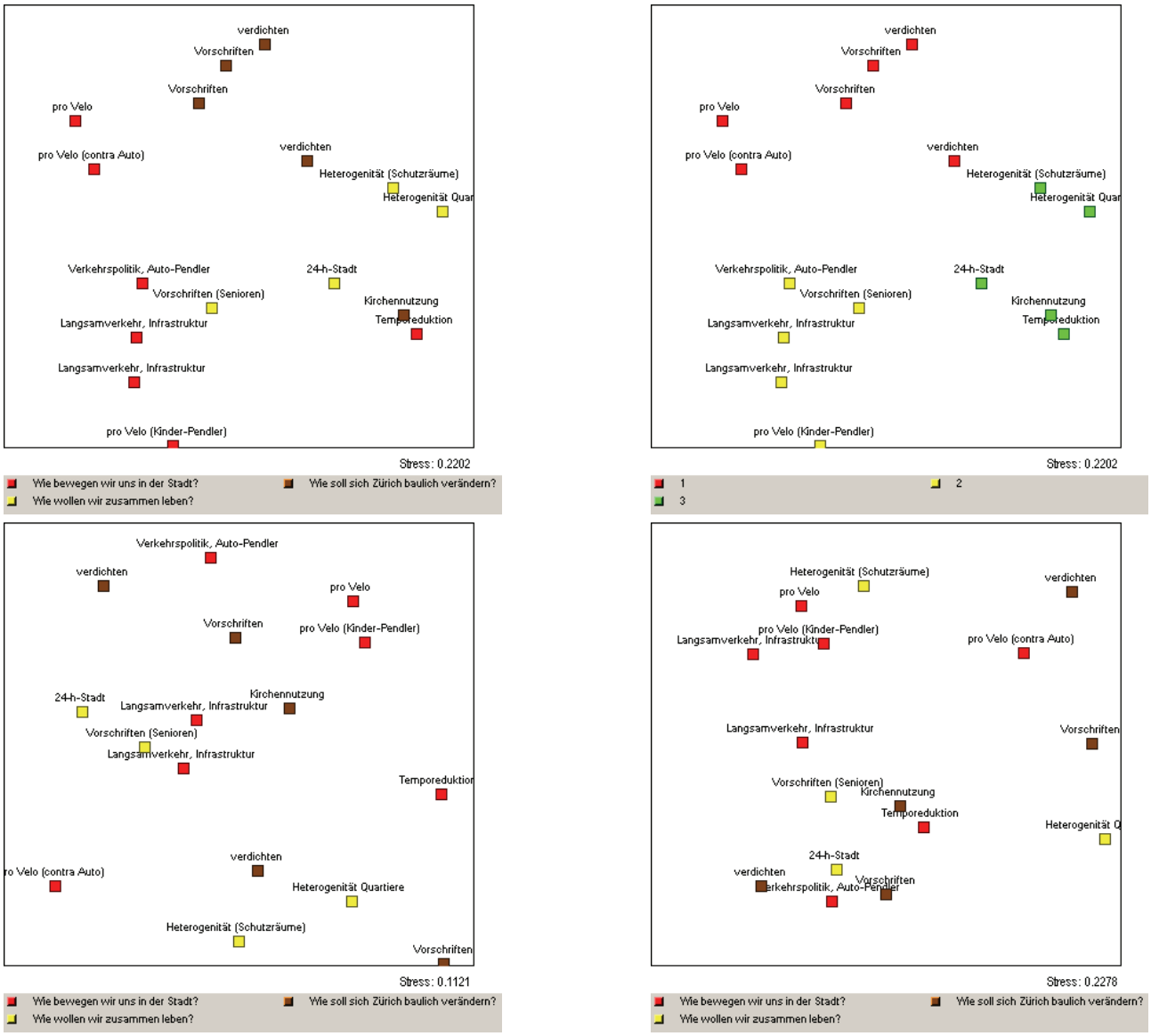


Abbildung 53: Links oben ist die Karte der 16 Texte, wenn der ÜK paarweise berechnet wird, eingefärbt sind die Foren. Rechts oben ist dieselbe Karte, jedoch sind die Items nach der Cluster-Ids der HCA eingefärbt. Links unten wurden die ersten 20, rechts die ersten 50 TargetWords gemäss Wortfrequenzenmethode gerechnet.

Die Wortfrequenzenmethode wird nun angewendet und die 20 (resp. 50) Wörter mit den höchsten Frequenzabweichungen als TargetWords bestimmt, dann erfolgt die Behofung und Skalierung. Beide resultierenden Karten sind in Abbildung 53 (unten links/rechts) ersichtlich.

Beide Strukturen sind inhaltlich nicht «schön». Vor allem bei der Karte mit 50 TargetWords fällt die Struktur komplett auseinander. Bei 20 TargetWords ist knapp die Struktur aus dem ÜK erkennbar, jedoch

sind die Verschiebungen der Texte inhaltlich nicht nachvollziehbar. Es kann also bereits an dieser Stelle gesagt werden, dass die Wortfrequenzenmethode für diese kleine Datenbasis nicht geeignet ist.

14.3.5 Vergleich mit dem Ansatz der häufigsten 20 Wörtern

Im Gegensatz dazu funktioniert der Ansatz der häufigsten Wörtern ganz gut (s. auch Abb. 56). Es wurden die 20 häufigsten Wörter der 16 Texte bestimmt und als Keywords genommen. In Abbildung 54 ist links die resultierende Karte, rechts prokrustet mit der eben erwähnten Karte von Abbildung 56 (links). Die Strukturen sind relativ ähnlich, gerade wenn man bedenkt, auf welch unterschiedliche Weise die zugrunde liegenden Keywords berechnet wurden.

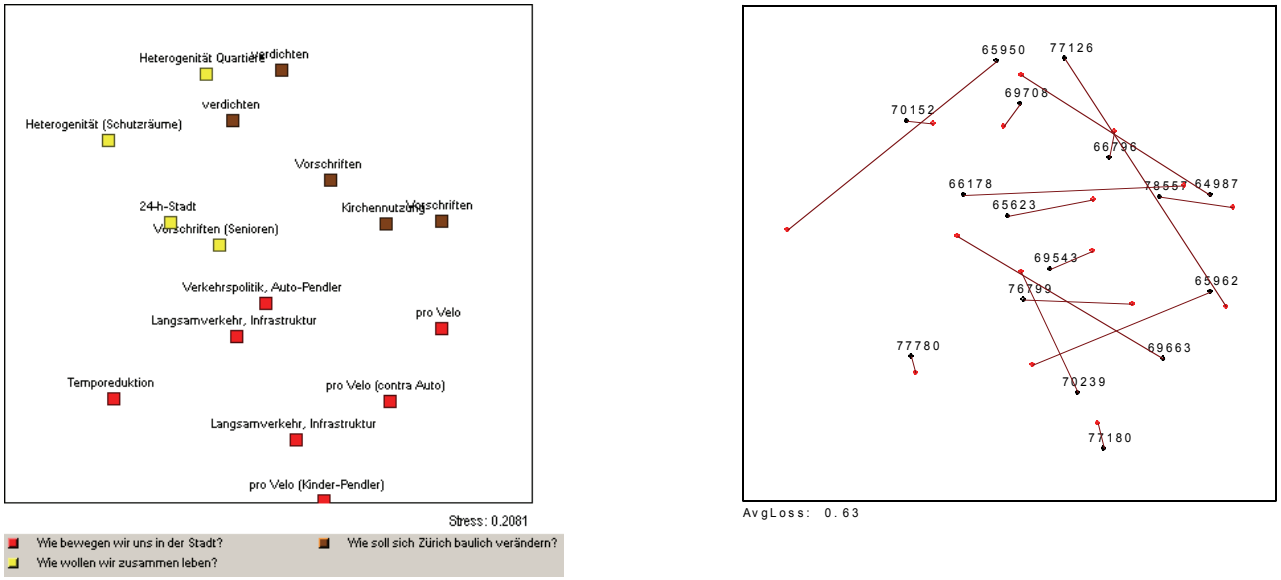


Abbildung 54: Links wurden die häufigsten 20 Wörter der 16 Texte als Keywords bestimmt und rechts mit der Karte prokrustet, die mit der Wortfrequenzenmethode (über den gesamten Textkorpus) berechnet wurde (schwarz: die 20 häufigsten Wörter, rot: Wortfrequenzmethode).

14.4 Resultate

14.4.1 Karte basierend auf dem Überlappungskoeffizienten

Es zeigt sich ein ÜK-typisches Bild (Abb. 55): Die Karte ist relativ stark geclustert und semantisch grösstenteils sinnvoll, jedoch auch «oberflächlich» in dem Sinn, dass Texte nah beieinander platziert werden, die zwar Wörter gemeinsam haben, semantisch aber nicht sehr ähnlich sind (Item 70152 und

65950). Auch die Items 66178, 78557 und 77780 haben inhaltlich keine Gemeinsamkeiten, wurden aber in einen eigenen Cluster platziert.

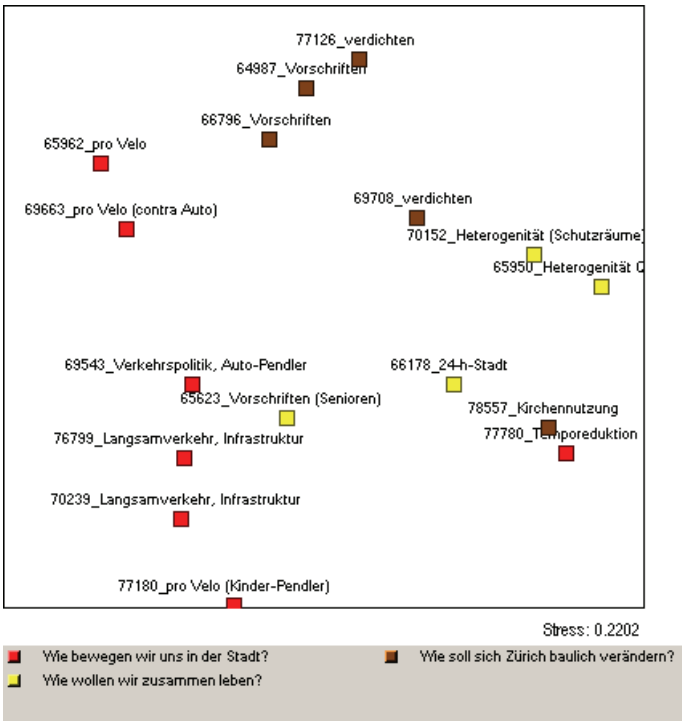


Abbildung 55: Karte basierend auf dem Überlappungs-koeffizienten

14.4.2 Häufigste Wörter als Keywords (Datenbasis: gesamter Textkorpus)

Alle Wörter des Textkorpus wurden ausgezählt und die 50, resp. 250 häufigsten Wörter als Keywords genommen. In der Abbildung 56 sind diese beiden Karten ersichtlich. Die linke Karte, basierend auf den 50 häufigsten Wörtern, ist recht ähnlich zu derjenigen, die mit dem ÜK berechnet wurde, jedoch sind einzelne Texte semantisch sinnvoller platziert. So ist Item 78557 (Kirchennutzung) näher bei den Verdichtungs-texten und Item 77780 (Temporeduktion) ist in der Nähe des Verkehrsclusters. Die rechte Karte, basierend auf den 250 häufigsten Wörtern, fällt inhaltlich auseinander: Durch die Aufnahme von weniger häufigen Wörtern scheint ein Rauschen hinzu zu kommen, das das spärliche Signal der strukturgebenden Keywords übertönt. Dieser Umstand wird, wie man im nächsten Kapitel lesen kann, auch bei der Wortfrequenzmethode auftreten.

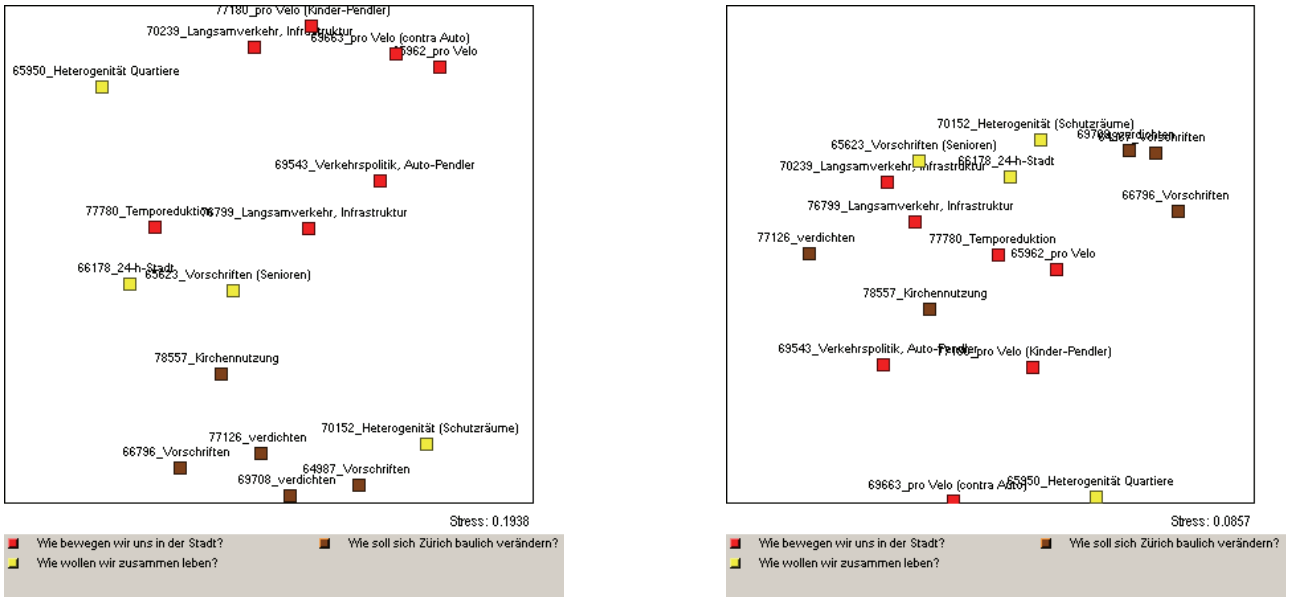


Abbildung 56: Links wurden als Keywords die 50 häufigsten Wörter gewählt, rechts die 250 häufigsten Wörter. Interessanterweise scheint die tiefere Anzahl von Keywords eine bessere semantische Struktur zu liefern.

14.4.3 Wortfrequenzmethode (Datenbasis: gesamter Textkorpus)

Die Keywords wurden gemäss Kap. 14.3 berechnet und die 16 Texte mit diesen 190 Keywords behoft (s. Abb. 57, rechts). Die Struktur der resultierenden Karte ist komplett sinnfrei. Nimmt man jedoch nur die 50 Wörter mit der höchsten Frequenzabweichung, sieht das Bild ganz anders aus (Abb. 57, links).

Abbildung 57: Keywords aus Wortfrequenzmethode – links wurden die ersten 50 Keywords genommen, rechts 190.

Die semantische Strukturierung ist recht gelungen. Das Verkehrscluster ist beisammen, mitsamt dem Item 77780 (Temporeduktion). Item 78557 (Kirchennutzung) ist in der Nähe der Verdichtungsitems und die drei Vorschriften-Items sind ebenfalls benachbart.

14.4.4 Instabilität und Plastizität der Karten

Durch die Art und Weise, wie die TargetWords in einem Text bestimmt werden, kommt eine Zufallskomponente in die Berechnung der Karten: Wenn mehr TargetWords als eine festgelegte Anzahl (20) in einem Text gefunden werden oder der Koeffizient der TargetWords über einem bestimmten Mass (0.14) liegt, wird per Zufall eine Auswahl aus den Fundstellen gezogen, damit diese Bedingungen erfüllt sind.

Zudem sind die Karten alle nicht sehr stabil und deshalb mit Vorsicht zu geniessen. Die gezeigten Karten stehen für Tendenzen der jeweiligen Berechnungsart. Es kann durchaus sein, dass in bestimmten Fällen der ÜK semantisch sinnvoller strukturierte Karten produziert als die Wortfrequenzenmethode. Folgendes Split-Half-Beispiel verdeutlicht die relative Instabilität.

Aus der Wortfrequenzenmethode wurden die ersten 100 Wörter extrahiert und der Reihe nach nummeriert. Die Wörter mit den ungeraden Ziffern kamen in eine erste Gruppe und die geraden in eine zweite Gruppe. Diese zwei Gruppen stellten zwei Keyword-Listen dar. Die obige Auswahl von 16 Texten wurde danach behoft und zwei Karten berechnet, die mittels Prokrustes-Transformation aufeinander gelegt wurden (s. Abb. 58).

Man sieht, wie sich die Karte zwar deutlich ändert, jedoch nicht komplett anders wird. Ein Item springt, zwei weitere ändern ihre Positionen markant, der Rest verschiebt sich zwar, jedoch bleibt die Makro-strukur erhalten.

Trotz dieser relativen Instabilität lassen sich Aussagen über die verschiedenen Berechnungsverfahren machen; sie sind nur nicht haarscharf zu trennen.

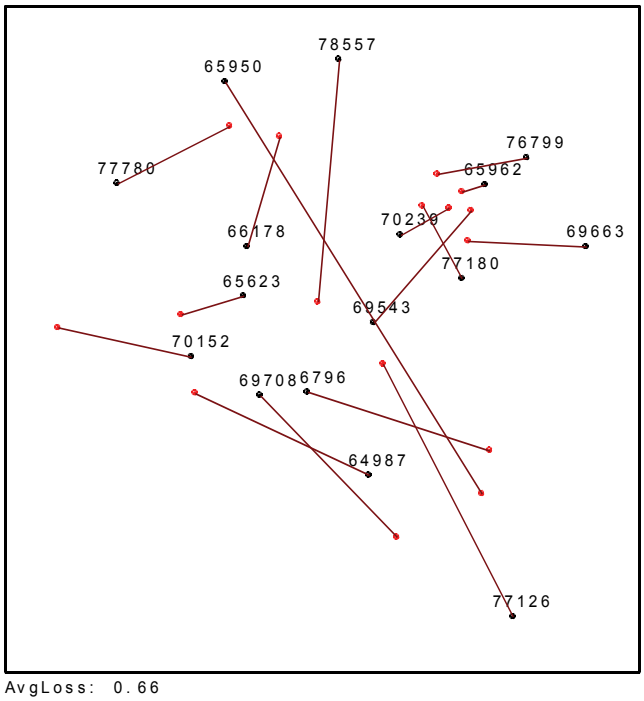


Abbildung 58: Prokrustes-Transformation der beiden Split-Half-Karten

14.5 Diskussion

Die Wortfrequenzmethode und das Verfahren der häufigsten Keywords funktionieren beide recht gut und produzieren bei einer grossen Textbasis ähnliche Ergebnisse. Bei beiden muss beachtet werden, dass nur eine gewisse Anzahl Keywords genommen wird. Warum das so ist und was die optimale Anzahl von Keywords ist, wird hier nicht untersucht. Ebenso wenig wird untersucht, wie sich die beiden Verfahren gegenüber manuellen Methoden behaupten können.

Bei einer kleinen Textbasis ist das Verfahren der häufigsten Keywords verlässlicher. Welcher der beiden Methoden der Vorzug zu geben ist, kann aufgrund der getätigten Berechnungen nicht gesagt werden. Es ist jedoch zu erwarten, dass die Wortfrequenzmethode grundsätzlich stabiler wird, wenn die Datenbasis deutliche semantische Strukturierungen aufweist.

15 KeywordII-Analyse

15.1 Überblick

Die KeywordII-Analyse ist ein von uns entwickeltes Verfahren, mit dem bestehende Keywordlisten auf den aktuellen Datenbestand hin optimiert werden können. Nur diejenigen Keywords bleiben in der Liste bestehen, welche eine grösstmögliche Trennung von Bedeutungsclustern ermöglichen. Das Verfahren kann halb- oder vollautomatisiert genutzt werden.

15.2 Einleitung

Wie schon in Kapitel 14 (Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient) gezeigt, ist die Auswahl der TargetWords ein wichtiges Teil im Räderwerk der Hofmethode. Mit der Wortfrequenzmethode wurde ein Verfahren gefunden, mit dem sich diese Arbeit automatisieren lässt. Im vorliegenden Kapitel wird ein alternatives Verfahren gezeigt, das einerseits zur Bestimmung von Keywords eingesetzt, andererseits aber dazu benutzt werden kann, bestehende Keywordlisten zu straffen. Vor allem bei vorwiegend statischer Datengrundlage können so Berechnungsvorgänge optimiert werden.

Die KeywordII-Analyse ist ein zweistufiges Verfahren; sie ist auf eine vorgängige Kategorisierung angewiesen. Diese kann aus den Metadaten stammen oder aus einer Clusteranalyse. Im zweiten Fall ist die KeywordII-Analyse vollautomatisiert. In diesem Kapitel wird jedoch nur der erste Fall getestet.

15.3 Vorgehen

15.3.1 Konzept der KeywordII-Analyse

Grundidee

«Gute» Keywords sind solche, die die vorliegende Datenbasis auf semantisch differenzierte Weise strukturieren. Dazu sollten sie in möglichst vielen Texten vorkommen, gleichzeitig jedoch eine breite

konnotative Bedeutung besitzen. Die konnotative Bedeutung müsste sich in den Hofwörtern niederschlagen. Also sind «gute» Keywords solche, die in der Datenbasis weit verbreitet sind *und* Hofwörter haben, die möglichst nur in einem der distinkten Bedeutungscluster vorkommen.

Da stellt sich sofort die Frage, wie die Bedeutungscluster zustande kommen. Da eine grobe Einteilung in semantische Richtungen genügt (die Feinarbeit geschieht danach durch die KeywordII-Analyse, Behofung und NMDS), ist ein pragmatischer Ansatz eine allenfalls vorhandene Kategorisierung in den Metadaten zu verwenden.

Eine weitere Möglichkeit ist nicht mal auf diese Vorarbeit angewiesen, sondern lässt sich vollständig automatisieren: Die Daten werden erst aufgrund der häufigsten Stichwörter behoft und verglichen, anschliessend wird entweder eine Clusteranalyse durchgeführt oder es wird eine semantische Karte gerechnet und die Cluster aufgrund der Karte berechnet. In beiden Fällen dient die Clusterzugehörigkeit als Indikator für die semantische Kategorie.

Umsetzung

In einem ersten Schritt werden die häufigsten Wörter einer Datenbasis bestimmt. Mit einer gewissen Anzahl davon (das kann eine fixe Zahl sein oder sie wird in Abhängigkeit zu den damit gefundenen TargetWords (TargetWord-Koeffizienten) gesetzt) werden die Höfe bestimmt. Von allen Hofwörtern wird in einer Tabelle festgehalten, wie oft und in welchen Kategorien (alternativ: Clustern) sie vorkommen. Für jedes Hofwort wird der Diskriminierungsfaktor (das Verhältnis Intra-Extra-Cluster) berechnet: Anzahl Vorkommen innerhalb des Clusters dividiert durch die Totalanzahl. Da ein Hofwort in verschiedenen Höfen vorkommen kann, wird der durchschnittliche Wert ausgerechnet und nach Anzahl Vorkommen gewichtet.

Die Gewichtungsformel lautet:

$$Gew.\,Diskr.wert = \frac{Diskr.wert}{1.01 - Diskr.wert} * Anz.\,Vorkommen\,Hofwort$$

Die resultierende Hofwortliste wird nach dem Diskriminierungsfaktor sortiert. Die Texte werden wieder nach dieser Liste behoft, wobei nur so viele Keywords aus der Liste ausgewählt werden, bis bestimmte Rahmenvorgaben (Mindestabdeckung der Texte mit mind. einem TargetWord und TargetWords-Koeffizient) erfüllt sind.

15.3.2 Methode

Das Verfahren wird anhand der edulap-Datenbank (s. Kap. 23, Projekt edulap) getestet, da hier eine besonders umfangreiche und domänenspezifische Keywordliste mit knapp 500 Einträgen vorhanden ist, die einen hohen Qualitätsstandard setzt. Aus den insgesamt über 300 Texten werden alle deutschsprachigen Texte (186) herausgezogen, sowie deren häufigste Stichwörter (46). Von 60 zufällig ausgewählten Texten werden die TargetWords mitsamt den Höfen bestimmt. Anhand der in den Metadaten zugewiesenen psychologischen Bereichskategorien wird die KeywordII-Analyse mit einem TargetWord-Koeffizienten von mind. 0.08 durchgeführt, zusätzlich mit der Vorgabe, dass alle Texte mindestens ein TargetWord[21] enthalten müssen. Mit nur 24 Keywords sind beide Bedingungen erfüllt. Die Texte werden mit den neuen Keywords behoft und die semantische Karte berechnet.

Diese Karte wird schliesslich mit einer Karte prokrustet, die auf den TargetWords der psychologischen Keywordliste beruht.

15.4 Resultate

In der Tabelle 6 sind die 46 häufigsten Keywords aufgelistet. Nachdem die 60 Texte damit behoft und der KeywordII-Analyse unterzogen wurden, resultieren die 24 Keywords, welche in der Tabelle 7 aufgelistet sind und in der linken Tabelle zusätzlich fett markiert sind. Man erkennt, dass die meisten Keywords aus dem oberen Bereich der linken Tabelle stammen. Eine inhaltliche Aussage ist schwierig und würde für dieses kleine Sample zu weit gehen. So ist es beispielsweise nachvollziehbar, dass ähnliche Wörter wie «betroffene» und «betroffenen», «störung» und «störungen» auf die gleiche Weise als KWII markiert (oder eben nicht) werden, jedoch nicht plausibel, dass «experiment» und «experimentellen» zwar beide KWII-Wörter sind, nicht aber «experimente».

21 Das wurde in diesem Fall auf diese Art gemacht, damit später die Prokrustes-Transformation durchgeführt werden konnte – diese funktioniert nur, wenn sich in beiden Karten dieselben Items befinden. Normalerweise setzt man eine Untergrenze von ca. 95% der Texte, die mind. ein TargetWord enthalten müssen. Einige Texte dürfen durch die Maschen fallen, weil es meist sog. Exoten gibt, die einen ungewöhnlichen (oder kurzen) Inhalt haben und die Keywordliste überproportional verlängern würden. Hätten wir das hier auch so gemacht, wäre ein Text (Id 3116, «Farbe») unbehoft geblieben und die Keywordliste hätte sogar nur 19 Einträge umfasst.

Tabellen 6/7: Die linke Tabelle zeigt die 46 häufigsten Wörter der 60 Texte. Fett gedruckt sind diejenigen Keywords, welche durch die KWII-Analyse (rechte Tabelle) ermittelt wurden.

1	**symptome**
2	**psychologie**
3	**störung**
4	studierenden
5	**störungen**
6	**betroffenen**
7	**daten**
8	**oft**
9	**verfahren**
10	entwicklung
11	**verhalten**
12	ergebnisse
13	**zwei**
14	**dabei**
15	**vorlesung**
16	**experimentellen**
17	experimente
18	stunden
19	**angst**
20	mindestens
21	behandelt
22	**experiment**
23	**betroffene**
24	methoden
25	**grundlagen**
26	**menschen**
27	einführung
28	hypothesen
29	wichtigsten
30	theorien
31	expra
32	**test**
33	sowie
34	wahrnehmung
35	**aufgaben**
36	**soziale**
37	**verschiedenen**
38	**geht**
39	themen
40	literatur
41	häufig
42	**meist**
43	kriterien
44	anhand
45	psychologischen
46	veranstaltung

1	betroffenen
2	symptome
3	experimentellen
4	verfahren
5	störung
6	zwei
7	oft
8	meist
9	aufgaben
10	störungen
11	vorlesung
12	daten
13	betroffene
14	verhalten
15	soziale
16	test
17	verschiedenen
18	psychologie
19	grundlagen
20	experiment
21	angst
22	dabei
23	geht
24	menschen

In Abbildung 59 ist links die Karte zu sehen, welche basierend auf den 46 häufigsten Keywords erstellt wurde. Die Karte wird vom Text 3136 dominiert, der nur ein einziges TargetWord besitzt, welches keine Ähnlichkeiten zum Rest aufweist. Ansonsten ist die Struktur recht gelungen: Die klinischen Texte bilden einen engen Bereich und die statistischen Texte und Forschungsmethoden sind beisammen. Sogar die ziemlich heterogen Texte der Bereiche Wahrnehmung und Denken sind benachbart.

Die rechte Karte zeigt die Struktur, welche auf den Daten der KeywordII-Analyse beruht. Mit nur 24 Keywords wurde eine äusserst ähnliche Karte produziert, die eine hohe inhaltliche Qualität aufweist.

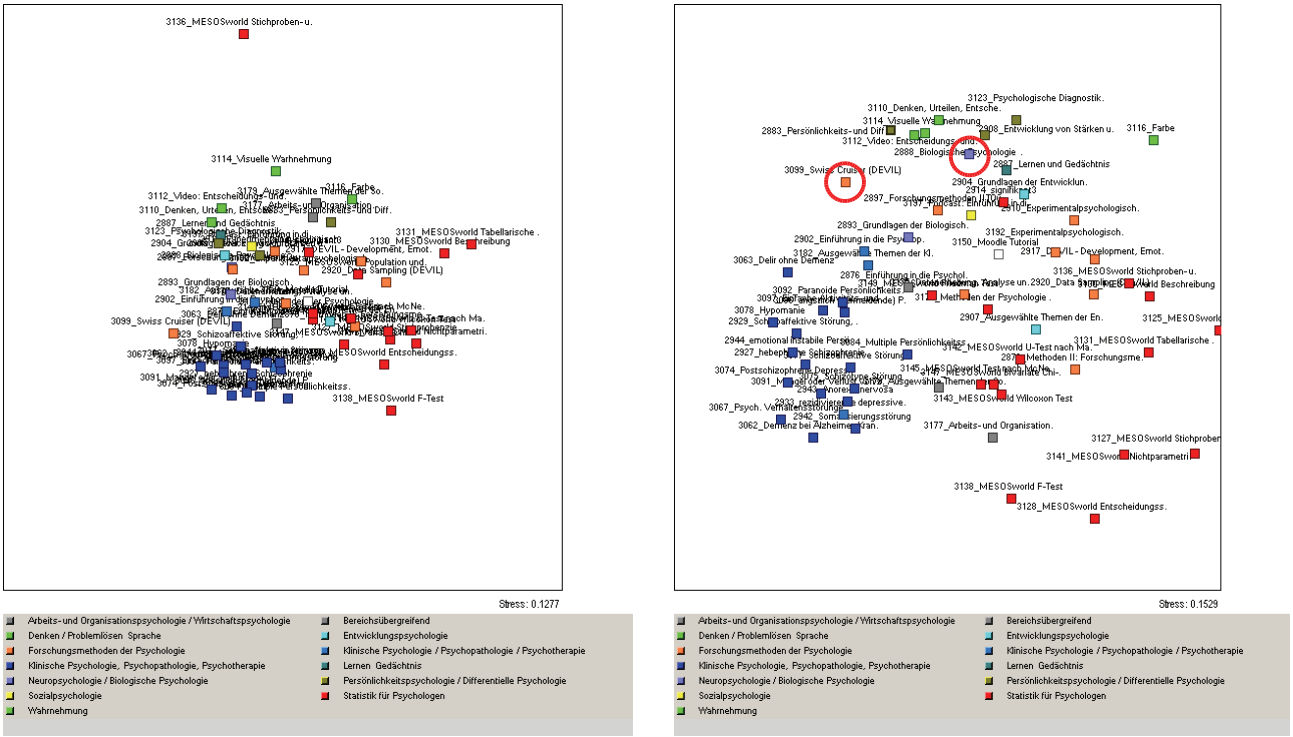
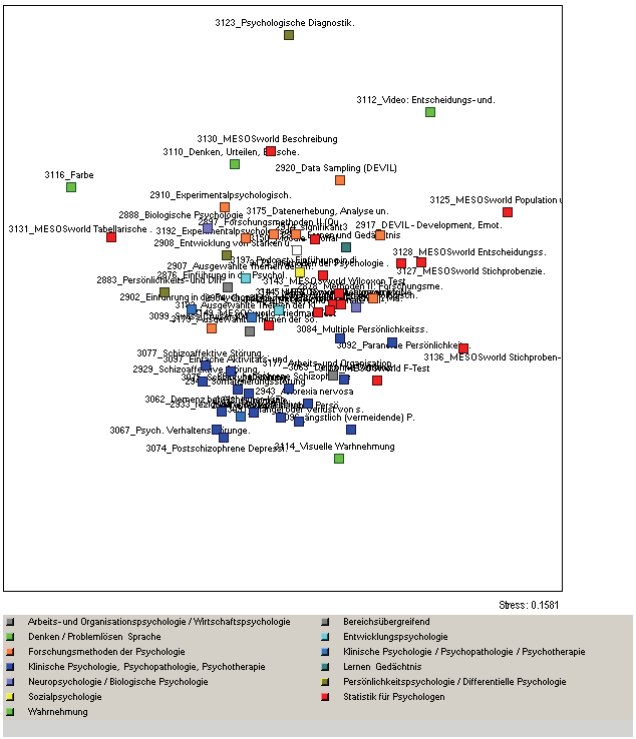


Abbildung 59: Links wurden die häufigsten 20 Wörter der 16 Texte als Keywords bestimmt und rechts mit der Karte prokrustet, die mit der Wortfrequenzenmethode (über den gesamten Textkorpus) berechnet wurde (schwarz: die 20 häufigsten Wörter, rot: Wortfrequenzmethode).

So befindet sich zwischen dem klinischen und dem kognitiven Bereich das Item «3099 Swiss Cruiser», welches ein Spiel zum Testen des prospektiven Gedächtnisses beschreibt. Die klinischen Items sind an diesem Rand die inhaltlich allgemein gefassten (Grundlagen und Einführung). Auch das Item «2888 Biologische Psychologie» passt sehr gut in den kognitiven Bereich; besser als in den klinischen.

In Abbildung 60 ist zum Vergleich die Karte dargestellt, welche mit den TargetWords der umfassenden, psychologischen Keywordliste erstellt wurde. Die Strukturierung ist weniger ausgeprägt, als in der KWII-Karte. Besonders die Texte des kognitiven Bereichs liegen weit auseinander. Dennoch sollte man sich nicht täuschen lassen: Im Grossen und Ganzen ist die Grundstruktur ähnlich, was auch die Prokrustes-Transformation rechts zeigt. Einige Springer im peripheren Bereich machen die Karte unansehnlich, der innere Bereich ändert sich jedoch nur wenig.



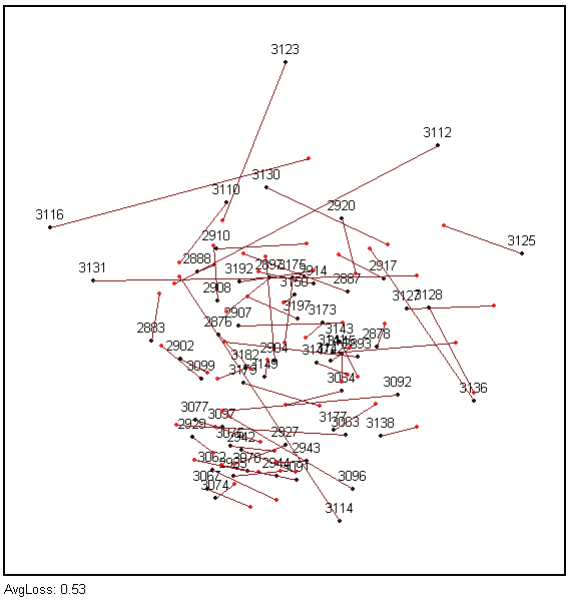


Abbildung 60: Links ist die Karte basierend auf der umfassenden, psychologischen Keywordliste. Rechts wurde diese Karte mit derjenigen der KWII-Analyse prokrustet (schwarz: Karte der psych. Keywords; rot: KWII).

15.5 Diskussion

Die KeywordII-Analyse kann dazu benutzt werden, eine kompakte Keywordliste zu erstellen, die auf den aktuellen Datenbestand zugeschnitten ist und ein Maximum an Semantik herausholt. Durch dieses Zurechtstutzen büsst man aber Flexibilität ein: Neue Texte können ev. nicht mehr behoft werden, beziehungsweise die Keywordliste müsste erweitert werden.

Das Kapitel ist zur Demonstration der KeywordII-Analyse gedacht. Wie sie flexibilisiert werden kann, beziehungsweise wie sie sich auf neue Texte auswirkt, müsste weiter erforscht werden.

16 Tagcloud-Verfahren von Semager

16.1 Überblick

Das Tagcloud-Verfahren ist ein Service eines externen Anbieters, der die semantisch relevantesten Wörter aus einem übergebenen Text extrahiert. Diese können benutzt werden, um eine Keywordliste für die Hofmethode zusammenzustellen. Im Vergleich mit einer umfassenden, nach psychologischen Kriterien zusammengestellten Liste liefert das Tagcloud-Verfahren qualitativ ebenbürtige Resultate. Das Verfahren eignet sich besonders in Anwendungsszenarien, in denen man flexibel auf neuartigen Text reagieren möchte oder um bestehende Verfahren zu ergänzen.

16.2 Einleitung

In Ergänzung zu den beiden besprochenen Algorithmen (s. Kap. 14 und 15) zur Keyword-Bestimmung haben wir ein externes Verfahren getestet. Matthias Schneider stellt auf seiner Website semager.de das Tagcloud[22]-API zur Verfügung, das die «wichtigsten Begriffe eines übergebenen Textes»[23] errechnet.

Gemäss Beschreibung (Schneider, 2009) verwendet Schneider ein selbst entwickeltes Verfahren, dem Wortkookkurenzen zu Grunde liegen, die einen semantischen Raum in einem neuronalen Netz aufziehen. Von einem übergebenen Text werden die Wortverwandtschaften berechnet und diejenigen Wörter retourniert, die über einem (ebenfalls übergebenen) Schwellwert liegen. Diese Wörter umschreiben im Idealfall das Kernthema des Ausgangstextes.

In diesem Kapitel führen wir einen einfachen Vergleich des Tagcloud-Verfahrens mit der Wortfrequenz-methode durch.

22 Zum Zeitpunkt der Programmierung der Tagcloud-Implementierung wurde das Verfahren noch «Textcloud» genannt. Das GUI des SemanticMappers verwendet noch diese alte Bezeichnung.

23 http://www.semager.de/info/semantic-business.php

16.3 Vorgehen

Aus 40 zufällig ausgewählten Texten des edulap-Projekts (s. Kap. 23, Projekt edulap) wird eine Karte basierend auf den knapp 500 Wörtern der dort verwendeten Regensburger Verbundklassifikation (RVK-Liste) erstellt. Anschliessend werden dieselben Texte (jedoch nicht entrauscht) dem Tagcloud-API übergeben und mit den retournierten Wörter eine neue Keywordliste erstellt. Die Parameter für Tagcloud sind:

- max. 20 Wörter pro Text
- Threshold: 4
- mit Berücksichtigung der Titel

Die 60 Texte werden mit diesen Keywords behoft und eine neue Karte erstellt. Beide Karten werden schliesslich prokrustet, um allfällige Qualitätsunterschiede offen zu legen.

Beide Karten werden mit dem Überlappungskoeffizienten für Listen berechnet, jedoch ohne Kategorienkonstante.

16.3.1 Fliesstext vs. entrauschtem Text

Dem Tagcloud-API könnte man entweder die entrauschten Texte übergeben oder die Originaltexte. Ein Test der beiden Varianten zeigte, dass der Unterschied minim ist: Das Tagcloud-API retourniert grösstenteils dieselben Wörter (s. Anhang S. 243).

16.4 Resultate

16.4.1 RVK-Liste

Die Karte basierend auf der RVK-Liste (Abb. 61) ergibt ein durchzogenes Bild: Die klinischen Texte bilden einen relativ engen Cluster im oberen Bereich. Die Statistik-Texte sind gegenüber im unteren Bereich angeordnet. Dazwischen sind Texte der Wahrnehmungs-, Entwicklungs-, Sozial- und Arbeits-psychologie, wobei die Wahrnehmungstexte und diejenigen der Sozialpsychologie nicht benachbart sind. Im Falle der Wahrnehmungstexte ist das nicht verwunderlich, bestehen sie teilweise aus viel zu kurzem

Beschreibungstext, doch die beiden Sozialpsychologietexte sind genügend lang und qualitativ gut. Auch der Text 3138 (links neben dem Klinik-Cluster) ist ein typischer Text seiner Kategorie und gehörte in den unteren Bereich.

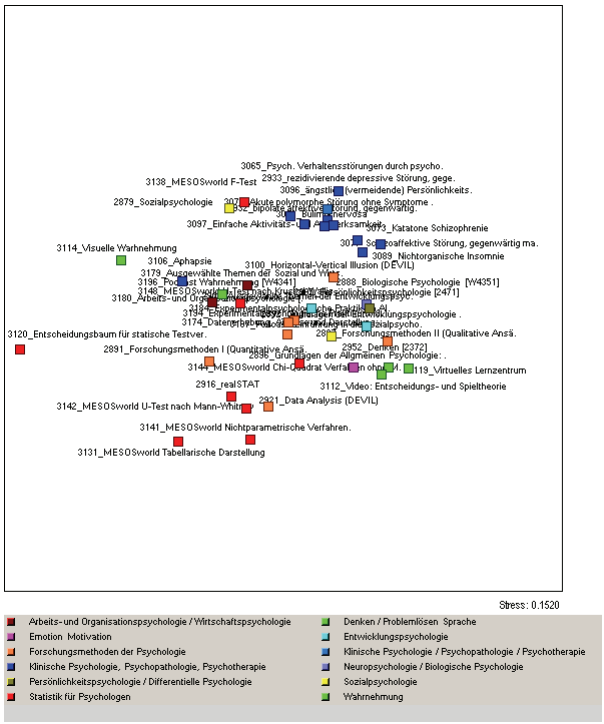


Abbildung 61: Karte basierend auf der RVK-Liste

16.4.2 Tagcloud

Das Tagcloud-Verfahren findet für die 40 Texte 62 Keywords. Ein Teil ist in Abbildung 62 wiedergegeben. Abbildung 63 zeigt exemplarisch die unterschiedlichen Verteilungen der TargetWords.

Die resultierende Karte (Abb. 64) weist folgende Charakteristiken auf: Die Klinik-Texte sind (ausser einem: ein Video über Aphasie, dessen Beschreibungstext nur Metainformationen über das Video selbst sind, keine Beschreibung der Störung) links in einem äusserst engen Bereich beisammen. Die restlichen Texte sind über den rechten Bereich verstreut, wobei die Texte der einzelnen Kategorien meist in der Nähe sind. Einzig die Forschungs- und die Statistiktexte sind nicht beisammen.

Abbildung 62: Ausschnitt aus der Tagcloud-Keywordliste

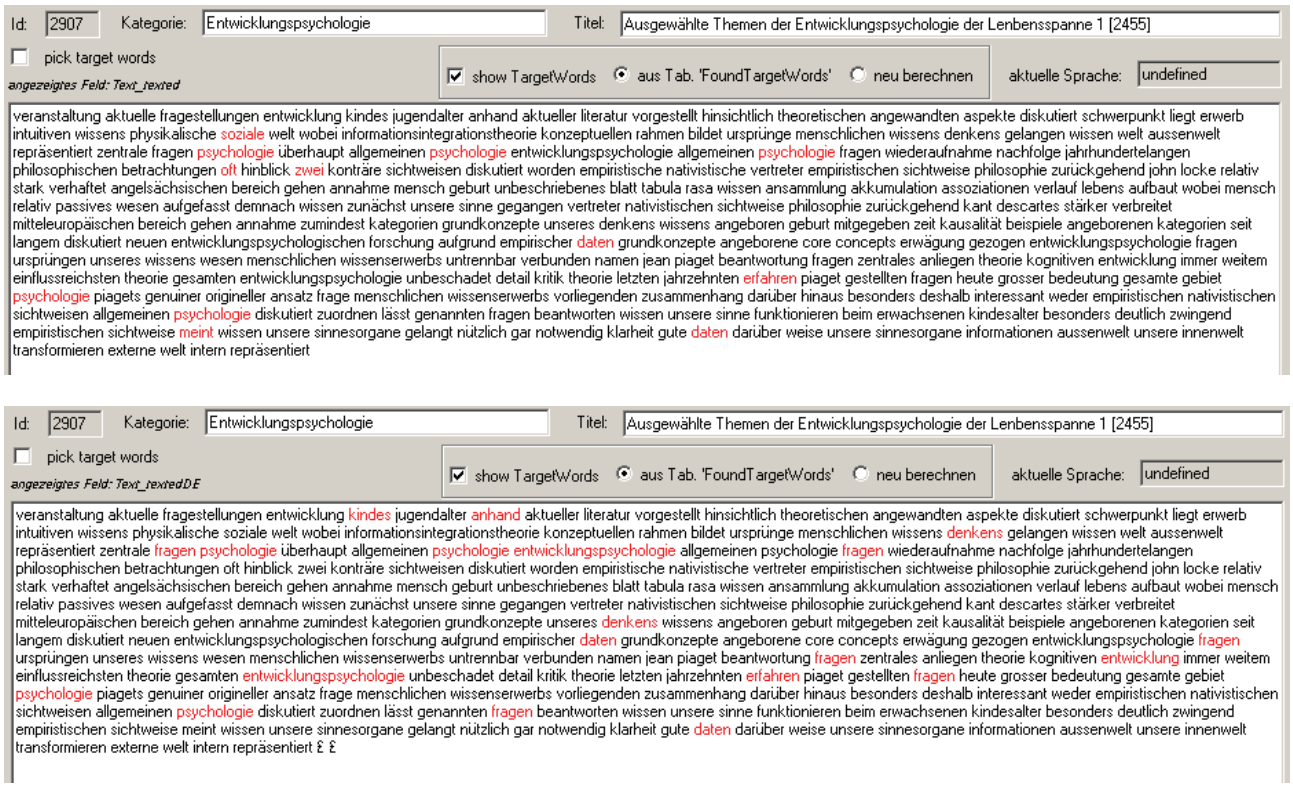


Abbildung 63: Oben: TargetWords der RVK-Liste; unten: TargetWords aus Tagcloud

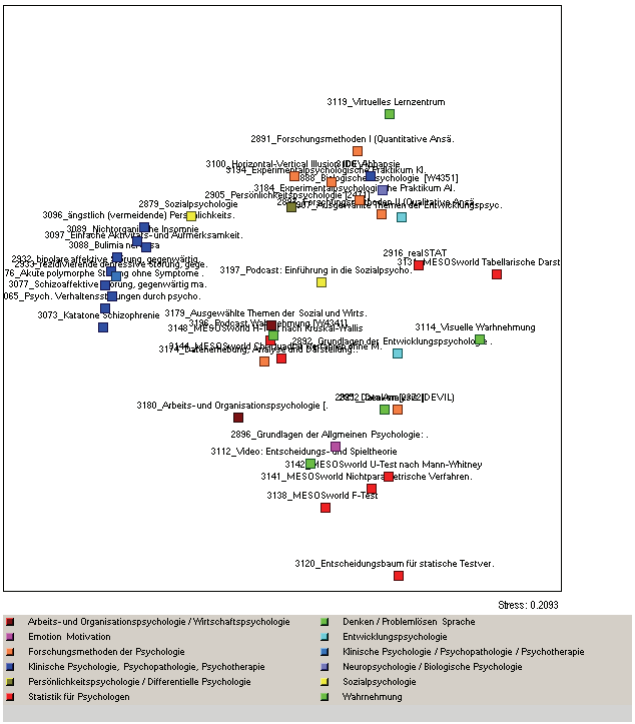


Abbildung 64: Karte basierend auf den TargetWords des Tagcloud-Verfahrens

Es macht zudem keinen grossen Unterschied, ob dem Tagcloud-API die entrauschten Texte oder die Originaltexte übergeben werden (s. Anhang 3).

16.4.3 Vergleich RVK-Liste und Tagcloud

In Abbildung 65 sind die Karten aus RVK und Tagcloud prokrustet. Die Makrostruktur ist in dem Sinne ähnlich, dass der klinische und der statistische Bereich das Feld aufspannen und sich dazwischen die restlichen Texte befinden.

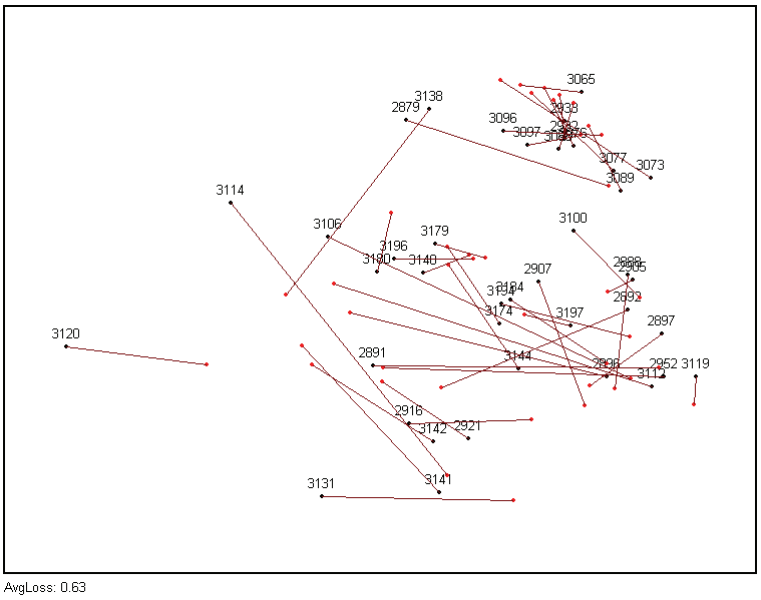


Abbildung 65: Prokrustes-Transformation der beiden Karten mit den TargetWords aus der RVK-Liste (schwarz) und Tagcloud (rot)

Die Anordnung dieser restlichen Texte ist zwar recht unterschiedlich, qualitativ jedoch nicht sehr verschieden. Beide Verfahren bilden eine Struktur, die von den klinischen und statistischen Texten dominiert wird, während sich die restlichen Items mit nur einer schwachen semantischen Struktur dazwischen aufreihen.

16.5 Diskussion

Das Tagcloud-Verfahren funktioniert mindestens so gut wie die vollumfängliche Liste aus dem psychologischen Katalog. Je nach Nutzungsszenario wird man auf einen fixen Katalog oder eine flexible

Lösung zurückgreifen wollen. Das Tagcloud-API hat den Vorteil, dass es auf die thematisch unterschiedlichsten Texte reagieren kann. Auf der anderen Seite macht man sich so von einem externen Dienstleister abhängig und man handelt sich einen Unsicherheitsfaktor bezüglich Serverkapazitäten ein.

Ein weitere Möglichkeit ist es, das Textcloud-Verfahren in Ergänzung zu anderen Verfahren einzusetzen. Beispielsweise kann per Wortfrequenzmethode eine Keywordliste erstellt werden. Bei Texten, bei denen damit zu wenige TargetWords gefunden werden, werden mittels Tagcloud weitere TargetWords identifiziert und in die Keywordliste aufgenommen.

17 Repräsentanten-Algorithmus

17.1 Überblick

Der Repräsentanten-Algorithmus reduziert die Items einer Karte auf eine gewünschte Anzahl von Zielitems unter Beibehaltung der semantischen Struktur. So wird eine ursprüngliche Struktur durch wenige Items repräsentiert.

17.2 Einleitung

Möchte man die Anzahl Items in einer Karte verringern (um die Übersichtlichkeit zu erhöhen oder den maschinellen oder menschlichen Bearbeitungsaufwand zu reduzieren), ohne jedoch die Struktur als solche stark zu verändern, muss man sogenannten Repräsentanten bestimmen.

Beispielsweise werden aus 100 Items 20 ausgewählt. Ideal wäre es, wenn eine erneute Ähnlichkeitsberechnung dieser 20 Items eine Struktur ergäbe, die relational identisch mit der Originalstruktur wäre. Die semantischen Bereiche blieben bestehen und Nachbarn blieben Nachbarn. Zudem sollte die Dichteverteilung proportional wiedergegeben werden: Wenn vorher 25 Items in einem engen Cluster waren, sollten es jetzt 5 sein.

In diesem Kapitel wird der von uns selbst entwickelte Repräsentanten-Algorithmus beschrieben und einer groben Plausibilitätsanalyse unterzogen. Ebenso wird gezeigt, wie sich dieser Algorithmus zwischen den bereits beschriebenen Cluster-Algorithmen (s. Kap. 3, Zielitems in der NMDS nach Verteilungen bestimmen) einreiht.

17.3 Vorgehen

17.3.1 Funktionsweise des Repräsentanten-Algorithmus

Damit ein Item als Repräsentant ausgewählt wird, muss es genügend Nachbarn haben. Als Nachbarn gelten diejenigen Items, die innerhalb eines bestimmten Radius liegen. Dazu wird für jedes Item die

Summendistanz zu den Nachbarn ermittelt. In einem iterativen Vorgehen wird dasjenige Item mit der höchsten Summendistanz als Repräsentant markiert. Das Verhältnis «Anzahl Gesamtitems» /«Anzahl geforderter Repräsentanten» gibt an, wie viele Nachbarn durch das gefundene Item repräsentiert werden und somit im nächsten Iterationsschritt nicht mehr zur Verfügung stehen. Die übrigen, weiter entfernten Nachbarn, können weiterhin als Repräsentanten ausgewählt werden, sofern sie ihrerseits genügend Nachbarn haben. Diese Schritte werden solange wiederholt, bis die gewünschte Anzahl Repräsentanten erreicht sind.

Die notwendige Anzahl Repräsentanten ist situationsabhängig und wird manuell an den Algorithmus übergeben. Die Bestimmung der Radiusgrösse kann automatisiert werden. Folgende Formel dient als Vorschlag; sie liefert brauchbare Resultate, kann aber sicherlich verbessert werden. Sie wurde empirisch erhoben, ist also nicht theoriegeleitet.

Der Radius wird – wie bei den Clusteralgorithmen – abhängig von der maximalen Itemdistanz ausgedrückt:

$$Radius = \frac{Max.Itemdistanz}{x}$$

Der Divisor x wird mit folgender Formel ermittelt:

$$x = Anz.Zielitems + \left(0.2 - \frac{Anz.Zielitems}{TotalItems}\right) \cdot 20$$

In Worten: Der Divisor entspricht ungefähr der gewünschten Anzahl Repräsentanten (=Zielitems), wobei eine Korrektur bewirkt, dass bei sehr wenigen Zielitems der Radius nicht allzu gross wird und bei vielen Zielitems nicht allzu klein. Folgende Tabelle zeigt, wie sich die Radiusgrösse bei unterschiedlicher Anzahl von Zielitems verändert.

Tabelle 8: Änderung des Divisors für die Radiusberechnung in Abhängigkeit der gewünschten Anzahl Zielitems (auf ganze Zahlen gerundet)

Gesamtanzahl Items	100	100	100	100	100	50
Anzahl Zielitems	50	20	10	5	1	20
Divisor	44	20	12	8	5	16

17.4 Methode

Das Verfahren wird anhand der edulap-Datenbank (s. Kap. 23, Projekt edulap) getestet. Es wird dasselbe Itemset wie in Kapitel 15 (KeywordII-Analyse) verwendet: 60 Texte werden zufällig aus dem Textkorpus gezogen, mit den 24 Keywords aus der KeywordII-Analyse behoft und die NMDS gerechnet. Die 60 Texte sollen durch 15 Repräsentanten vertreten werden. Diese Anzahl wird manuell festgelegt: Soll eine Grundgesamtheit durch nur einen Viertel ihrer Mitglieder vertreten werden, muss die Auswahl vorsichtig getroffen werden.

Der Divisor der Radiusgrösse wäre nach oben stehender Formel 14. Wir verwenden jedoch 8, um den Unterschied zum Prototypenalgorithmus besser demonstrieren zu können (s. Kap. 17.5.1).

17.5 Resultate

In Abbildung 66 ist links die ungefärbte Karte zu sehen. Die Texte sind relativ gleichmässig verteilt, wobei eine leicht dichtere Verteilung links und oben zu erkennen ist. Der rechte Bereich franst langsam aus. Der Algorithmus nimmt diese undeutlichen Charakteristiken sehr schön auf, wie in der Abbildung 66 rechts zu sehen ist: Die gefundenen Repräsentanten sind rot eingezeichnet. Die blauen Kreise indizieren die Radiusgrössen. Der linke Bereich ist deutlich vertreten, während im ausgefransten Bereich rechts kein Repräsentant gefunden wird.

In Abbildung 67 wurde die Karte mit den 15 Repräsentanten neu berechnet und mit der Ausgangskarte prokrustet, wobei nicht die 60 Items genommen wurden, sondern nur die Koordinaten der 15 Repräsentanten. Die Gesamtstrukturen stimmt erstaunlich gut überein und Nachbarn bleiben Nachbarn. Einzig das Item 3114 springt und das Item 3130 verschiebt sich in den Nachbarbereich.

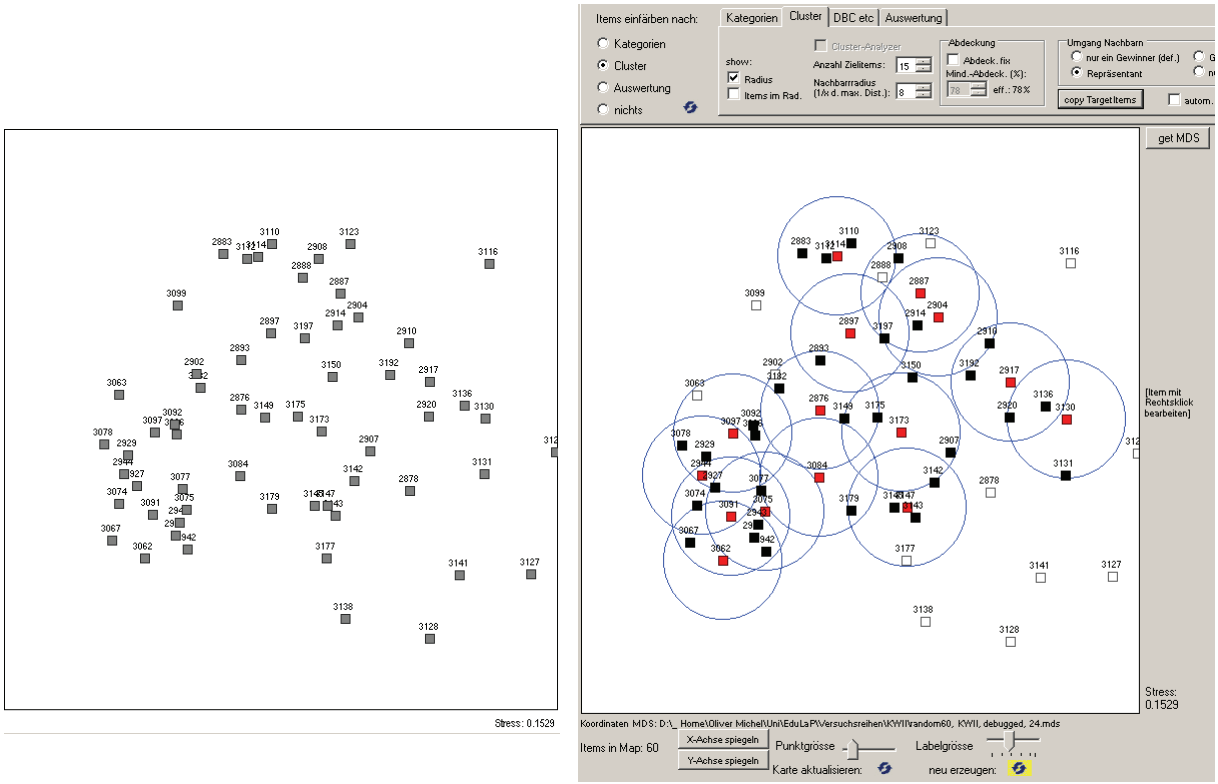


Abbildung 66: Links: die ungefärbte Karte mit den 60 Items; rechts: Screenshot der Karte mit den Repräsentanten und deren Radien.

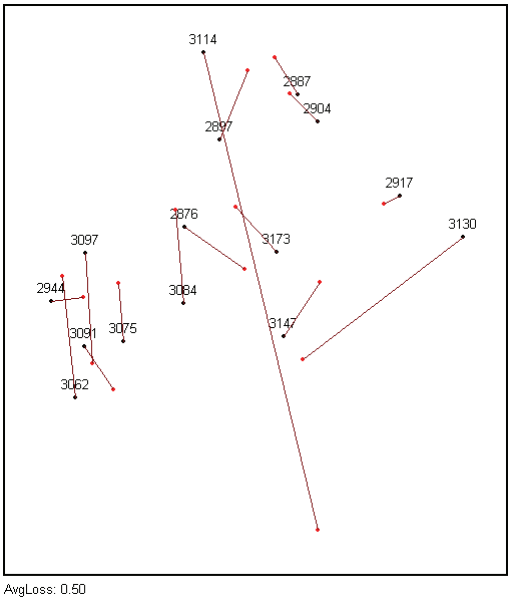


Abbildung 67: Prokrustes-Transformation der neu berechneten Repräsentantenkarte (rot) mit den Positionen der originalen Karte (schwarz)

17.5.1 Zunehmende Dezentralisierung

Im Anhang (Kap. 3) werden verschiedene Möglichkeiten zur Determinierung von Zielitems besprochen. Hier sollen die verschiedenen Auswirkungen der dort entwickelten Algorithmen im Zusammenhang mit dem Repräsentanten-Algorithmus dargestellt werden. Je nach Anforderung an die zu repräsentierende Zentralisierung lässt sich so ein bestimmter Algorithmus auswählen. Abbildung 68 zeigt immer dieselbe Karte, die auch in diesem Kapitel benutzt wurde. Als Zielvorgabe wurden 15 Zielitems und eine

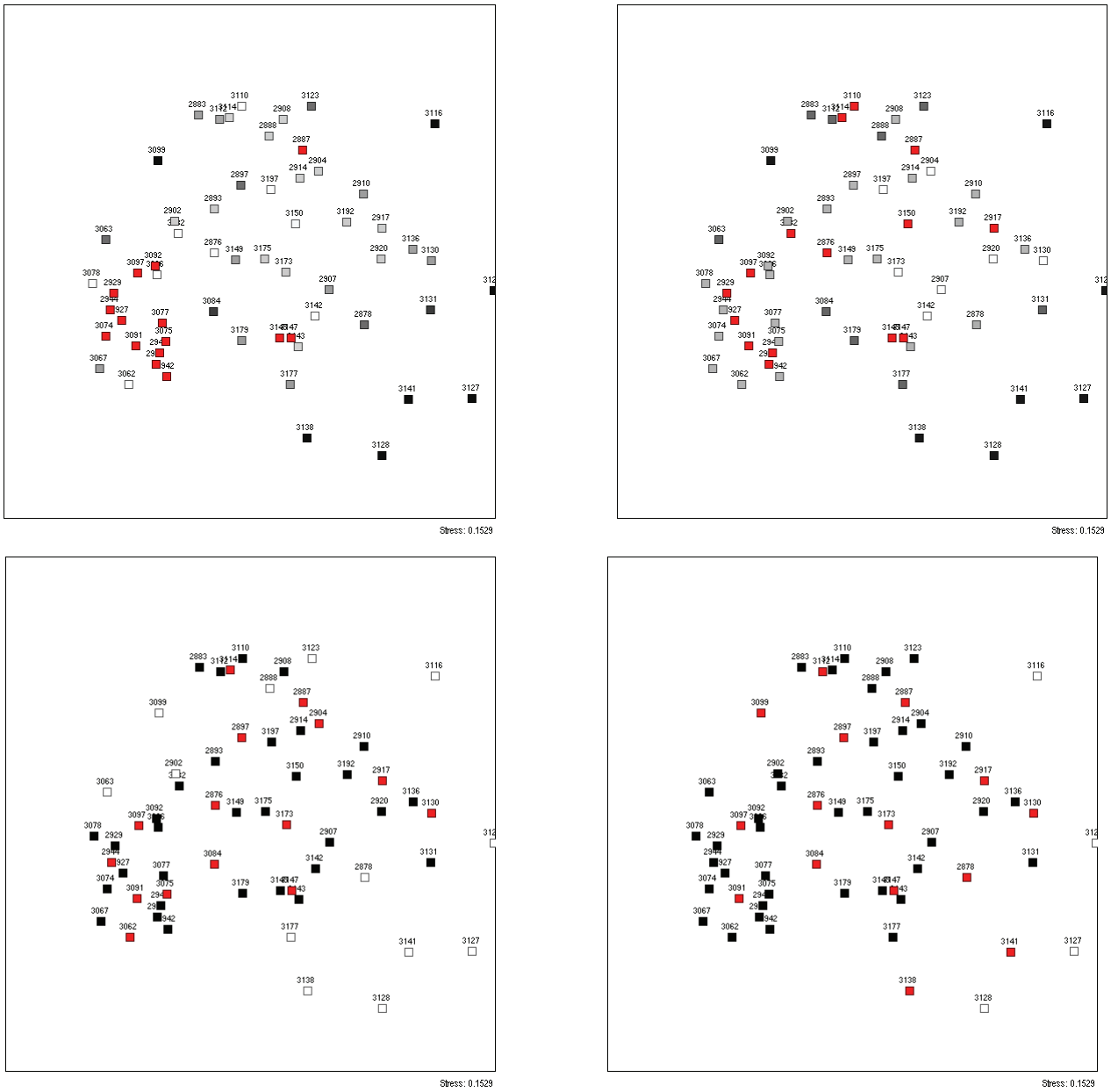


Abbildung 68: Verwendete Algorithmen im Uhrzeigersinn: «nur nach Werten», «Gewinner zählt nur einmal», «Repräsentant», «Zielitem eliminiert Nachbarn»

Radiusgrösse von 1/8 gewählt. Der Algorithmus «nur nach Werten» (oben links) berücksichtigt nur die zentralen Items; «Gewinner zählt nur einmal» (oben rechts) lockert diese Zentralisierung etwas auf; «Repräsentant» (unten links) wählt die Zielitems gemäss ihrer Dichte; «Zielitem eliminiert Nachbarn» (unten rechts) spannt das ganze Feld auf. Die Verteilung der Zielitems wird in dieser Reihenfolge also immer dezentraler.

17.5.2 Nachteil des sequenziellen Vorgehens

Durch das beschriebene iterative Verfahren zur Bestimmung desjenigen Items mit der höchsten Summendistanz lässt man sich auf einen Nachteil ein: Das in einem Iterationsschritt ermittelte Item ist nicht unbedingt optimal, wenn man auch den nächsten Iterationsschritt beachten würde.

Abbildung 69 verdeutlicht dies: Rot markiert ist der sichtbare Horizont. Ist dieser Radius gegeben, wäre das eingefärbte grüne Item das Gewinneritem (Karte links), weil es fünf Nachbarn abdeckt. Im nächsten Schritt wäre dann das Item links oberhalb davon das nächste Gewinneritem (Karte rechts), das noch drei neue Nachbarn abdeckt.

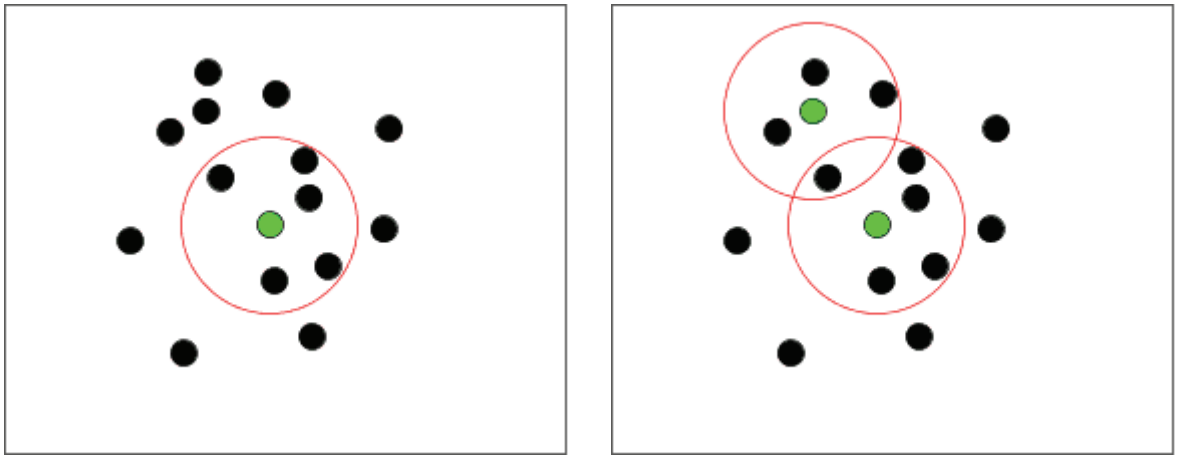


Abbildung 69: Im beschriebenen iterativen Verfahren ist die Auswahl der Gewinneritems nicht optimal.

Global betrachtet wäre es aber günstiger, wenn die beiden Gewinneritems so verteilt wären, wie in Abbildung 70: So würden insgesamt 9 Items abgedeckt.

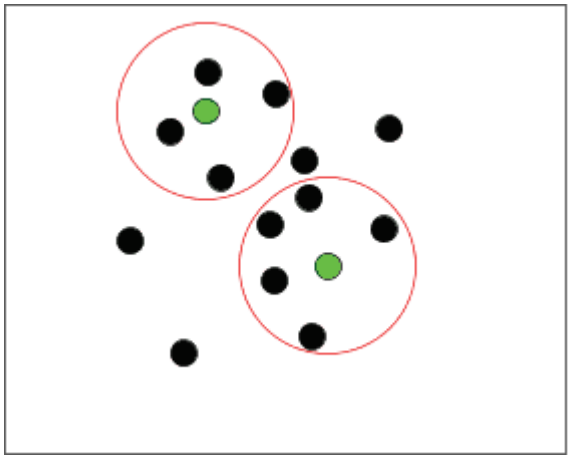


Abbildung 70: Eingezeichnet ist die hypothetische optimale Auswahl der Gewinneritems

Der Algorithmus müsste also den nächsten Iterationsschritt bereits in die aktuelle Berechnung miteinbeziehen.

Die Auswirkung des verwendeten einfachen iterativen Verfahrens ist allerdings marginal und wird in dieser Arbeit in Kauf genommen.

17.6 Diskussion

Der Repräsentanten-Algorithmus funktioniert sehr gut: Die Auswahl einer gewünschten Anzahl von Zielitems erfolgt so, wie angefordert: Die Dichteverteilung wird angemessen repräsentiert und eine anschliessende Neuberechnung der Auswahl ergibt eine Karte, die die Gesamtstruktur der Originalkarte recht gut wiedergibt. Das Ziel eine grosse Anzahl von Items auf eine überschaubarere Anzahl zu reduzieren, wird damit erreicht.

Im Unterschied zum Prototypen-Algorithmus (Kap. 3, Zielitems in der NMDS nach Verteilungen bestimmen) nimmt der Repräsentanten-Algorithmus mehr Rücksicht auf die Dichteverteilung einer Karte, während der erstgenannte schneller das Feld der Karte aufspannt.

18 Einbezug von Metadaten: Kategorienkonstante

18.1 Überblick

Falls eine a-priori-Kategorisierung vorhanden ist, kann diese Information genutzt werden, um falsche Platzierungen in semantischen Karten zu berichtigen. Das wird realisiert, indem der Textähnlichkeitswert um die Kategorienkonstante erweitert wird. Die Karte wird somit zusätzlich strukturiert, ohne dass die berechneten semantischen Relationen zerstört werden.

18.2 Einleitung

Das Kerneinsatzgebiet der Hofmethode (HM) sind unstrukturierte Textdaten. Falls dennoch Metadaten vorhanden sind, könnten diese in den Ähnlichkeitsvergleich miteinbezogen werden, um die Qualität der semantischen Strukturierung zu steigern. Gerade bei einer vorgängigen Kategorisierung ist diese Möglichkeit verlockend. Was auf den ersten Blick nur Vorteile hat, könnte sich auch negativ auswirken. Dann nämlich, wenn zwei Items die gleiche Kategorie zugeschrieben wurde, sie inhaltlich jedoch in andere Kategorien gehörten. Es ist ja gerade der Vorteil der HM, dass sie *nicht* auf eine vorgängige Kategorisierung angewiesen ist. Der Wert einer identischen Kategorisierung sollte also so klein gewählt werden, dass er nur berichtigend wirkt, falls ein Item beispielsweise in der Kartendarstellung springt oder gar keine Ähnlichkeiten aufweist und deshalb falsch platziert wird, nicht aber die durch die HM berechnete semantische Struktur übertönt.

In diesem Kapitel wird anhand des edulap-Datensatzes (s. Kap. 23, Projekt edulap) untersucht, wie sich die Berücksichtigung der psychologischen Inhaltskategorie mit verschiedenen Gewichtungen auf die semantische Strukturierung auswirkt. Die Allgemeinkategorie «bereichsübergreifend» wird zudem gesondert betrachtet.

18.3 Vorgehen

Es werden manuell drei Subsets aus den edulap-Daten zusammengestellt. Beim ersten Subset werden typische Texte aus den drei distinkten Kategorien «Wahrnehmung», «Statistik» und «Klinische Psychologie»

ausgewählt, wobei ein Text aus der klinischen Psychologie ein besonderer ist: Bei Text «3108 – Hiebe statt Liebe» handelt es sich um eine Lehr-DVD zur Erkennung von Kindsmisshandlung. Inhaltlich gehört das Thema eindeutig zur klinischen Psychologie, der Duktus ist allerdings ein methodischer. Rein inhaltlich gesehen ist das Item nicht eindeutig kategorisierbar; hierfür braucht es die Kenntnis des Curriculums in Form der Metadaten.

Als Baseline werden erst die Textähnlichkeiten berechnet, ohne dass die Kategorisierung berücksichtigt wird, anschliessend wird bei einer identischen Kategorisierung zweier Texte zum Textähnlichkeitswert eine Konstante addiert, die mit verschiedenen Werten variiert wird.

Die Formel für die Textähnlichkeitsberechnung wird somit um die Kategorienkonstante erweitert:

$$Ähnlichkeitswert_{(TextA:TextB)} = Aufsummierte\ Hofwerte + Gewicht_{ÜK} * Wert_{ÜK} + Kategorienkonstante$$

Beim zweiten Subset kommen die Texte ebenfalls aus drei Kategorien, nämlich «Klinische Psychologie», «Statistik» und «Forschungsmethoden», jedoch stammen die beiden letzteren aus einem ähnlichen Bereich der Psychologie. Hier geht es also um den Aspekt, dass zwei Kategorien zwar unterschiedlich bezeichnet werden, inhaltlich jedoch nah miteinander verwandt sind. Die Kategorienkonstante muss so bestimmt werden, dass sie keine Scheincluster bildet.

Beim dritten Subset schliesslich werden drei Texte unterschiedlicher Kategorien zum Subset 2 dazu genommen, wobei ein Text die Inhaltskategorie «bereichsübergreifend» hat (Item 2876, «Einführung in die Psychologie»). Diese Kategorie ist ein Sonderfall: Ein bereichsübergreifender Text hat potenziell mit allen anderen Bereichen etwas zu tun und muss deshalb gesondert behandelt werden. Die Platzierung eines solchen Textes würde man intuitiv in der Kartenmitte vermuten, wo der Abstand zu allen anderen Bereichen am geringsten ist. Wir berücksichtigen das, indem die Kategorienkonstante auf einen minimalen Wert gesetzt wird, wenn ein Item eines Vergleichspaares zur Kategorie «bereichsübergreifend» gehört, unabhängig davon, wie die Kategorie des Vergleichsitems lautet.

Bei allen Texten werden Listen mit einem ÜK-Gewicht von 1 gewichtet. Die Keywords stammen aus der RVK-Liste. Als Normierung wurde TotalTargetWords gewählt (s. Kap. 5, Normierung der Textähnlichkeitswerte: SharedTargetWords vs. TotalTargetWords).

18.4 Resultate

18.4.1 Kategorienkonstante

Subset 1

Das Subset 1 strukturiert auch ohne Berücksichtigung der Kategorienzugehörigkeit recht gut (s. Abb. 71). Alle Texte der Kategorie «Wahrnehmung» bilden einen eigenen Cluster. Die Texte der «Klinischen

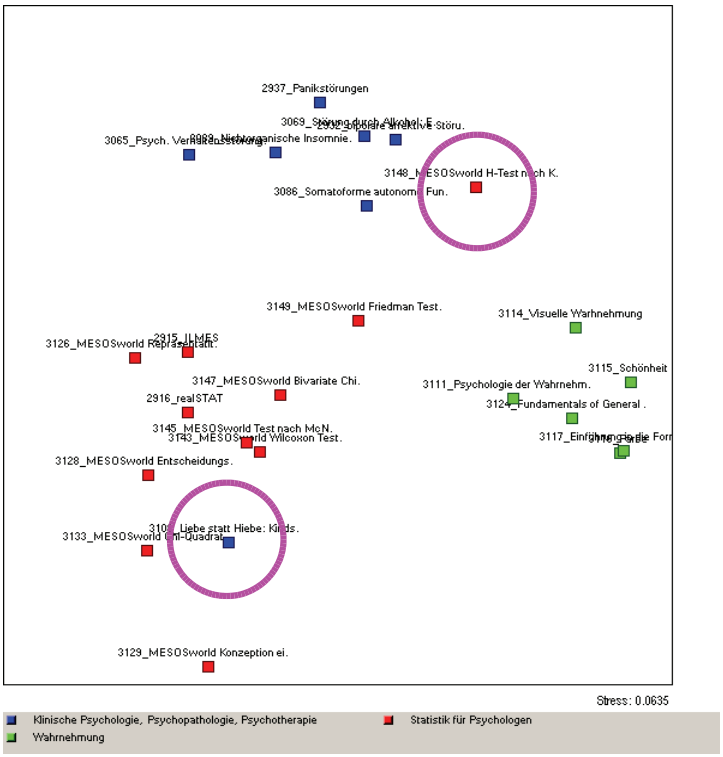


Abbildung 71: Subset 1, ohne Kategorienkonstante. Die Clusterung ist bis auf zwei Items klar ersichtlich.

Psychologie» bilden einen weiteren Cluster, in dessen Peripherie sich ein Statistik-Text befindet (Item 3148). Die Statistik-Texte liegen weniger eng beieinander, bilden trotzdem einen eigenen Bereich. Das exotische Item 3108 kommt tatsächlich in den Statistik-Bereich zu liegen – die Hofmethode scheint die diagnostischen Ähnlichkeiten stärker zu berücksichtigen, als das Thema der klinischen Verhaltensstörung durch Kindsmisshandlung.

Wird nun bei identischer Kategorisierung zweier Texte eine Konstante von 1 hinzu addiert, clustert die resultierende Karte äusserst stark (s. Abb. 72). Zu stark. Die weiche, semantische Struktur geht verloren und weicht einer dichotomen Kategorisierung, die innerhalb einer Kategorie nicht mehr interpretiert werden kann. Die Ähnlichkeitswerte werden fast vollständig durch die Kategorienwerte dominiert, was unerwünscht ist.

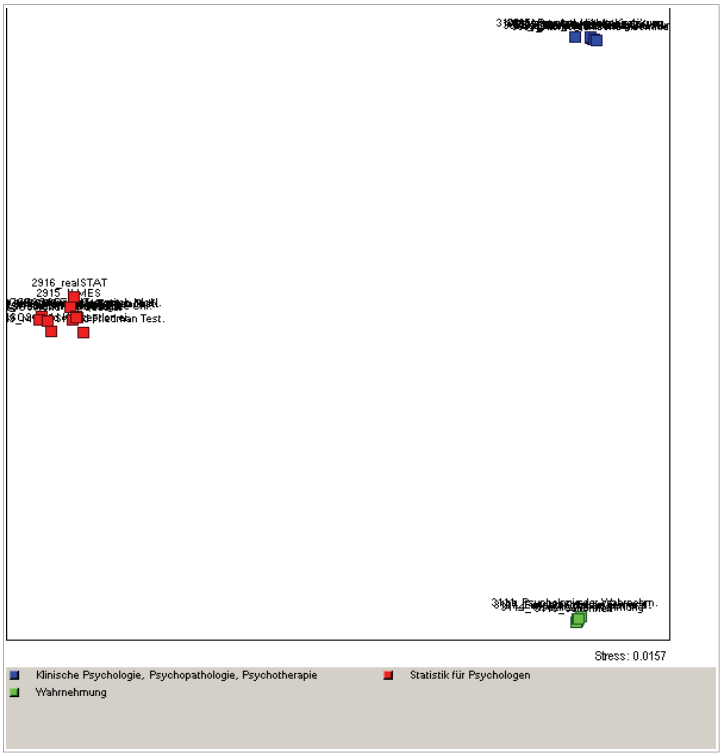


Abbildung 72: Subset 1, Kategorienkonstante von 1.

Auch bei einer Halbierung des Wertes auf 0.5 ist die Kategoriendominanz ungebrochen (Abb. 73), erst ab Werten kleiner als 0.1 beginnt die Intraclusterstruktur wieder interpretierbar zu werden (Abb. 74, 75). Da die Kategorisierungsinformation nur helfen soll, bestehende Fehler zu bereinigen, ansonsten aber möglichst ohne Wirkung bleiben sollte, ist der Wert möglichst klein zu wählen. Bei einem Wert von 0.02 ist eine gute Balance zwischen Kategorisierung und semantischer Strukturierung gefunden (Abb. 76): Die Fehler der Ausgangskarte sind behoben – die beiden Ausreisser befinden sich nun bei ihrem entsprechenden Cluster, dort aber am Rand.

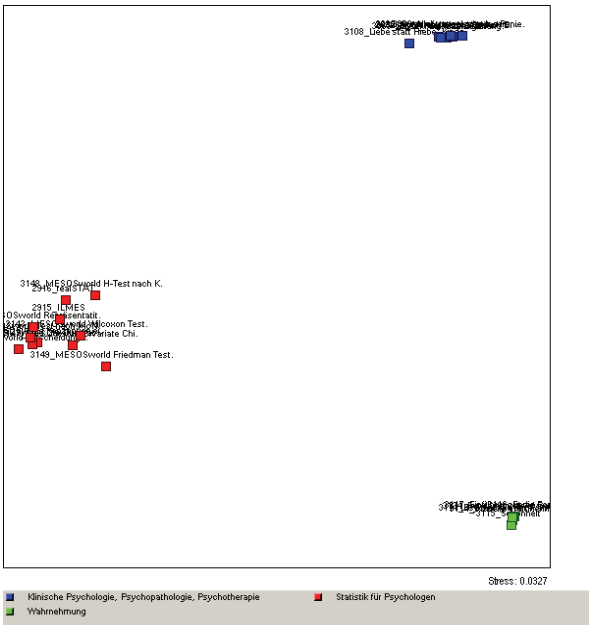


Abbildung 73: Subset 1, Kategorienkonstante 0.5

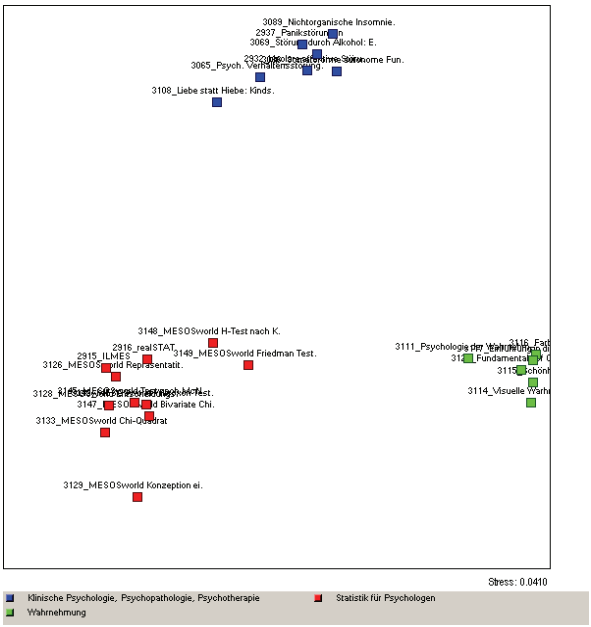


Abbildung 74: Subset 1, Kategorienkonstante 0.1

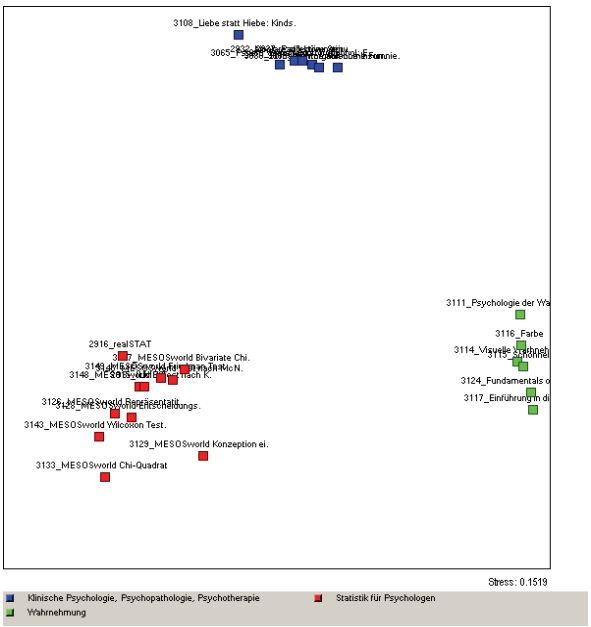


Abbildung 75: Subset 1, Kategorienkonstante 0.05

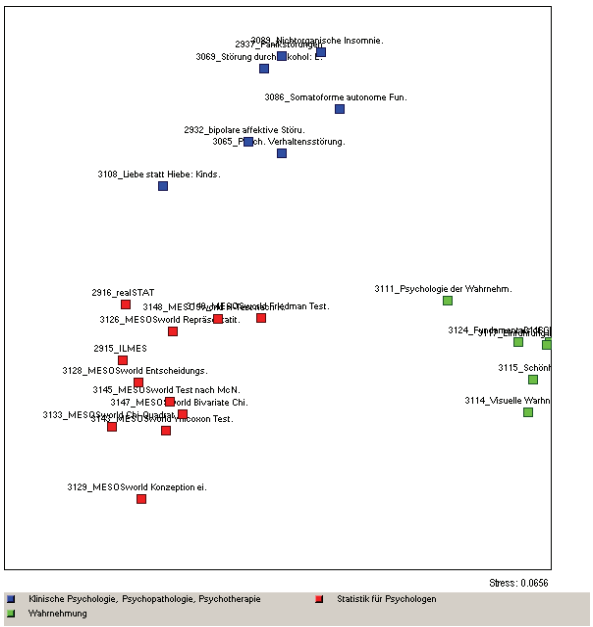


Abbildung 76: Subset 1, Kategorienkonstante 0.02

Subset 2

Subset 2 zeigt – ebenfalls ohne Kategorisierung – eine befriedigende Trennung der Kategorien (Abb. 78): In der linken Kartenhälfte sind die Statistik- und Forschungsmethodentexte, in der rechten die klinischen Texte. Die Durchmischung der beiden Kategorien «Forschungsmethoden» und «Statistik für Psychologen» ist semantisch gesehen durchaus nachvollziehbar, beim Anwendungsfall des edulap-Projekts

jedoch nicht ideal, zudem dürfte die Trennung zu den klinischen Texten klarer sein. Ein Kategorienwert von 0.02 hat genau den gewünschten Effekt (Abb. 77).

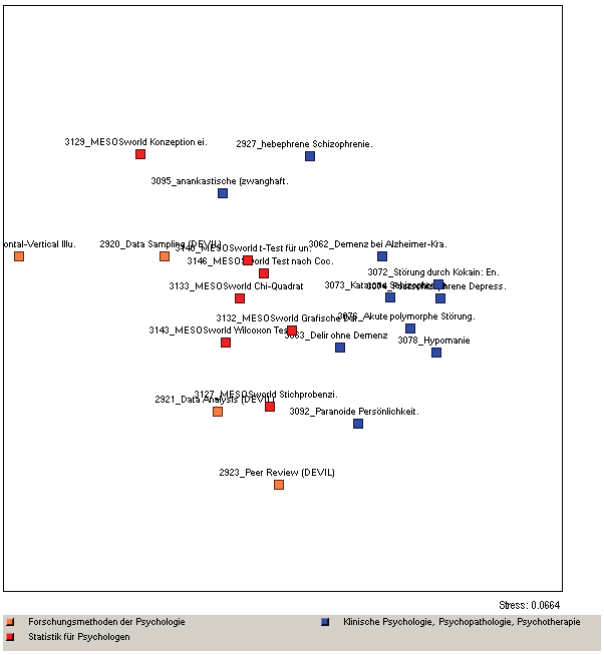


Abbildung 78: Subset 2, ohne Kategorien-konstante

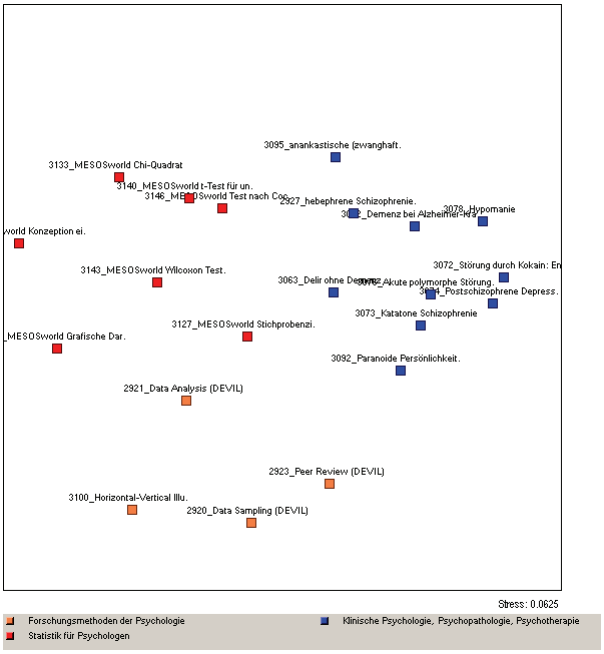


Abbildung 77: Subset 2, Kategorienkonstante 0.02

Der Blick auf einen Teil der Dreiecksmatrix zeigt, warum der Wert von 0.02 sinnvoll ist. Oft haben Items gar keine Ähnlichkeit, hier bewirkt ein Kategorisierungswert – sei er noch so klein – viel. Jedoch verfälscht ein kleiner Kategorisierungswert nicht eine bestehende Ähnlichkeit. Somit ist ein Wert von 0.02 als ideal anzusehen.

Tabelle 9: Subset 2, ohne Kategorisierungswert

	2920	2921	2923	2927	3062	3063	3072	3073	3074	3076	3078	3092	3095	3100	3127
2920															
2921	0														
2923	0	0.0769													
2927	0	0	0												
3062	0	0	0	0											
3063	0	0	0	0	0.0645										
3072	0	0	0	0	0.2528	0									
3073	0	0	0	0.0435	0.1429	0.0385	0.1306								
3074	0	0	0	0	0.2857	0	0.6439	0.2833							
3076	0	0	0	0	0.3595	0.0696	0.5607	0.3825	0.5586						
3078	0	0	0	0	0.0833	0	0.2902	0.0562	0.1667	0.2647					
3092	0	0	0.0833	0	0	0.0556	0	0.0879	0	0.0851	0.0909				
3095	0	0	0	0.0625	0.0714	0.0526	0	0	0	0	0	0			
3100	0	0	0	0	0	0	0	0	0	0	0	0	0		
3127	0	0.782	0	0	0	0.0588	0	0.0476	0	0.0385	0	0.0769	0	0	

Tabelle 10: Subset 2, Kategorisierungswert 0.02

	2920	2921	2923	2927	3062	3063	3072	3073	3074	3076	3078	3092	3095	3100	3127
2920															
2921	0.02														
2923	0.02	0.0969													
2927	0	0	0												
3062	0	0	0	0.02											
3063	0	0	0	0.02	0.0845										
3072	0	0	0	0.02	0.2728	0.02									
3073	0	0	0	0.0635	0.1629	0.0585	0.1506								
3074	0	0	0	0.02	0.3057	0.02	0.6639	0.3033							
3076	0	0	0	0.02	0.3795	0.0896	0.5807	0.4025	0.5786						
3078	0	0	0	0.02	0.1033	0.02	0.3102	0.0762	0.1867	0.2847					
3092	0	0	0.0833	0.02	0.02	0.0756	0.02	0.1079	0.02	0.1051	0.1109				
3095	0	0	0	0.0825	0.0914	0.0726	0.02	0.02	0.02	0.02	0.02	0.02			
3100	0.02	0.02	0.02	0	0	0	0	0	0	0	0	0	0		
3127	0	0.782	0	0	0	0.0588	0	0.0476	0	0.0385	0	0.0769	0	0	

18.4.2 Kategorie «bereichsübergreifend»

Zum Subset 2 wurden drei Texte mit unterschiedlichen Kategorien hinzugefügt: Diese separieren in einen eigenen Cluster, obwohl sie aus so verschiedenen Bereichen wie «Sozialpsychologie», «Kognitionspsychologie» und «bereichsübergreifend» stammen. Die Hofmethode findet keine starken, semantischen Ähnlichkeiten zu den restlichen Texten. Die vorherrschende Strukturierung der Karte beruht jedoch auf den Kategorien (die Berechnung wurde mit einer Kategorienkonstante von 0.02 durchgeführt). Da diese

drei Texte die Unähnlichkeit zu den restlichen Texten gemeinsam haben, platziert sie die NMDS nahe beieinander (s. Abb. 79 links).

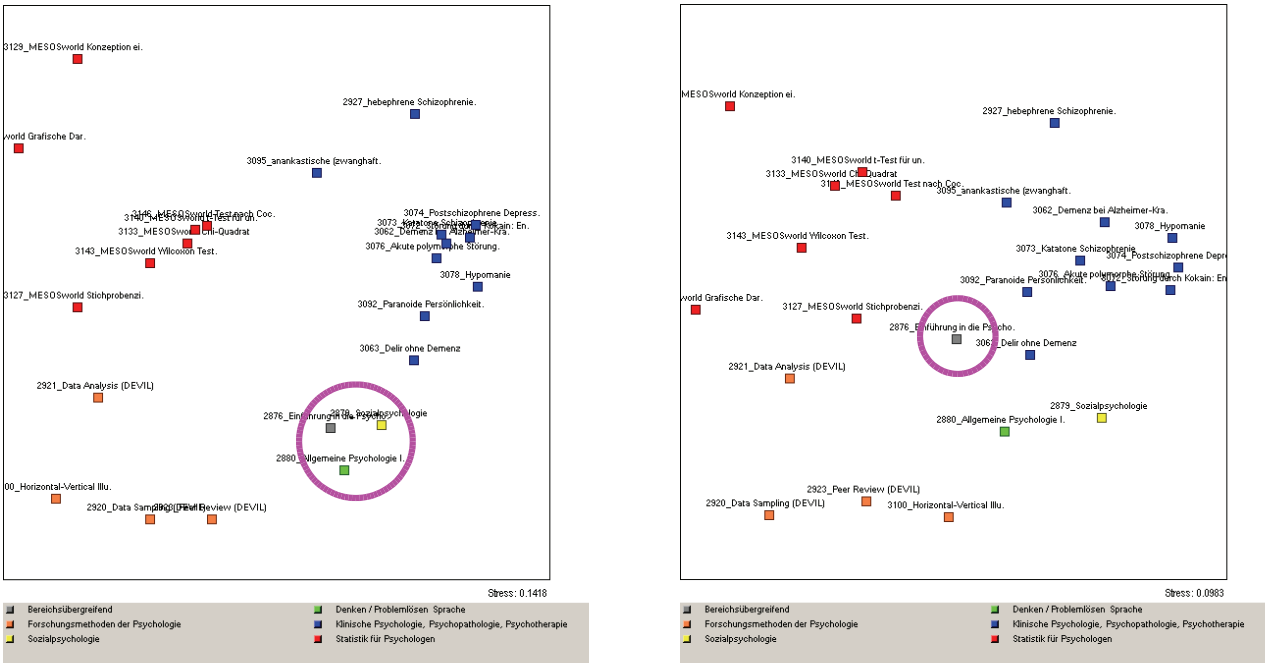


Abbildung 79: Links: Subset 3, Kategorienkonstante 0.02 – das allgemeine Item ist bei den beiden anderen Items mit uniquen Kategorien. Rechts: Subset 3, Kategorienkonstante 0.02, zusätzlicher Wert der allgemeinen Kategorie von 0.01 – das allgemeine Item rutscht korrekterweise in die Mitte.

Hier lässt sich der Effekt einer gesonderten Behandlung der allgemeinen Kategorie sehr anschaulich zeigen. Wird die Kategorienkonstante auf einen minimalen Wert von 0.01 gesetzt, wenn mindestens ein Item eines Vergleichspaares der Kategorie «bereichsübergreifend» angehört, rutscht das bereichsübergreifende Item 2876 weg von den beiden anderen Items mit uniquen Kategorien und wie gewünscht in die Kartenmitte (s. Abb. 79 rechts).

18.5 Diskussion

Ein kleiner Kategorienwert von 0.02 bringt Ordnung in eine Struktur, ohne bestehende semantische Ähnlichkeiten zu verfälschen. Falls also in einem Textkorpus diese Information vorhanden ist, sollte sie auch genutzt werden.

Gibt es im Textkorpus eine Kategorie, die Ähnlichkeiten zu allen anderen hat, wie «bereichsübergreifend», kann der Kategorienwert auf 0.01 gesetzt werden, wenn mindestens ein Item eines

Vergleichspaares diese Kategorie aufweist. Das Item wird dann in die Kartenmitte rutschen, wo es einen minimalen Abstand zu allen anderen Items aufweist.

Der ideale Kategorienwert ist abhängig von den Werten in der DEM, beziehungsweise deren Normierung. Anstatt einer Konstante wäre man flexibler, wenn der Kategorienwert dynamisch aus der DEM berechnet würde: Der Wert sollte zwischen dem Null-Wert und den kleinsten Ähnlichkeiten liegen.

19 Texte mit Listen: Einbezug des Überlappungskoeffizienten

19.1 Überblick

Listen eignen sich nicht für die Hofmethode, da sie keinen verwertbaren Kontext aufweisen. Kommen in einem Textkorpus Texte mit Listen vor, sollten diese mittels Überlappungskoeffizienten verglichen werden. Dadurch sinkt der Stress von semantische Karten.

19.2 Einleitung

Die HM nutzt zur Bedeutungsgenerierung den Kontext von Stichwörtern. Somit braucht die HM Fliesstext: Text mit redundantem Textmaterial, das Stichwörter in einer Wolke von Assoziationen einbettet. Diese Voraussetzung ist bei Stichwortlisten nicht gegeben. Die Einträge bestehen nur aus einzelnen Worten. Die Ordnung der Liste ist für die HM kaum nutzbar; ist die Liste beispielsweise nach einem chronologischen Datum oder nach Alphabet geordnet, kann die HM keine Semantik daraus ableiten. Die Hofwörter sind bei Listen meist arbiträr und ändern komplett, wenn die Liste nach einem anderen Kriterium umgeordnet wird.

Für die spezielle Textart «Stichwortliste» eignet sich ein Verfahren wie der Überlappungskoeffizient (ÜK) viel besser: Die Ordnung der Wörter spielt keine Rolle mehr. Zudem haben es in der Praxis Listen oft an sich, dass sie aus einem eingeschränkten, spezifischen Vokabular bestehen, das dem ÜK entgegenkommt.

Falls in einem Textkorpus Fliesstext und Listen gemischt auftreten, sollten diese beiden Textarten getrennt behandelt werden: Die HM sollte beim Fliesstext nach TargetWords suchen und der ÜK bei Listen zum Einsatz kommen. Anschliessend werden diese beiden Werte bei einem Textvergleich miteinander verrechnet: Die Textähnlichkeit setzt sich aus dem üblichen Hofvergleich der Fliesstexte und dem ÜK der Listen zusammen. Allerdings läuft man so Gefahr, dass Textähnlichkeiten aufgrund der Textart nicht erkannt werden: Was passiert, wenn ein Text nur aus Fliesstext besteht, der andere aber ähnlichen Inhalt in Form einer Liste hat?

Wir lösen das Problem, indem wir die Hofmethode einfach ebenfalls über die Listen laufen lassen, dort aber die Hofgrösse im Falle eines Fundes auf 0 setzen, weil der Kontext – wie beschrieben – irrelevant ist. Besteht nun in einem Textvergleich Text A nur aus Fliesstext und Text B nur aus einer Liste, kann der Vergleich trotzdem durchgeführt werden. Falls beide Texte aus Fliesstext und Listen bestehen, findet der Hofvergleich über die gesamten Texte mit variabler Hofgrösse statt, zusätzlich wird für die beiden Listen der ÜK berechnet und mit dem Hofvergleich verrechnet Damit nicht beispielsweise zwei kurze, unähnliche Listen die Hofähnlichkeiten unnötig herunterreissen, geht der Anteil der Listen an den Gesamttexten in die Berechnung ein:

$$\frac{(WortanzahlListenTextA + WortanzahlListenTextB)}{GesamtwortanzahlTextA + GesamtwortanzahlTextB}$$

Beispiel: Text 1 hat eine Fliesstextlänge von 200 Wörtern und eine Listenlänge von 100 Wörtern, Text 2 hat 250 Wörter Fliesstext und 10 Wörter in Listen. Der Anteil der Listen beträgt somit: (100 + 10) / (300 + 260) = 0.2. Der Ähnlichkeitswert aus dem Hofvergleich wird mit diesem Wert multipliziert.

Zusätzlich führen wir ein Gewicht ein, das den Einfluss des ÜK modifiziert.

Die Formel für den Textvergleich wird also mit dem ÜK-Wert, dessen Textanteil und Gewicht erweitert:

$$Ähnlichkeitswert_{(TextA\,TextB)} = Hofwert + Gewicht_{ÜK} * Wert_{ÜK} * AnteilListe + Kategorienkonstante$$

Die Kategorienkonstante (s. Kap. 18) ist in der Formel zwar angegeben, wird in diesem Kapitel aber nicht berücksichtigt, damit sich der ÜK-Effekt möglichst isoliert manifestieren kann.

In diesem Kapitel untersuchen wir, wie sich die gesonderte Behandlung von Listen auf die Textvergleiche auswirkt.

19.3 Vorgehen

Wir gewichten den ÜK mit verschiedenen Werten. Als Textmaterial wählten wir Texte des edulap-Projekts (s. Kap. 23, Projekt edulap), die mit Listen unterschiedlichster Längen durchsetzt sind. Beim Erfassen der Datenbank wurden Listen mit £-Zeichen umschlossen, damit sie maschinell erkennbar sind.

Beispiel Id 2878, «Methoden II: Forschungsmethoden und Statistik II»:

> Dieses Modul vermittelt ihnen weitere statistischen Verfahren, welche für die Auswertung psychologischer oder allgemein sozialwissenschaftlicher Daten herangezogen werden. Dabei ist zuerst ein gutes Verständnis des methodologischen und mathematischen Hintergrundes notwendig. Daraus kann abgeleitet werden, welchen Voraussetzungen die Daten zur Anwendung der verschiedenen statistischen Verfahren genügen müssen. Oder vom Standpunkt der praktischen Anwendung aus gesehen: Basierend auf den Eigenschaften der Daten und der zu prüfenden Hypothesen muss der korrekte und von der Teststärke her optimale statistische Test ausgewählt werden. Dieser Kurs behandelt Verfahren zur Überprüfung der zentralen Tendenz für Ordinaldaten (nicht-parametrische Verfahren), zum Vergleich von Varianzen, Vergleiche von Häufigkeitsverteilungen, regressive und korrelative Verfahren, Verfahren zur Anwendung bei wiederholter Messung (abhängige Daten) und mehrfaktoriellen Analysen. Dazu kommen Post-hoc-Analysen wie Einzelvergleiche und Kontrastverfahren zwecks Interpretation von Interaktionen, festen und zufälligen Effekten.
>
> £
>
> – Verfahren zur Überprüfung der zentralen Tendenz für Ordinaldaten (nicht-parametrische Verfahren): U-Test, Wilcoxon-Test, Vorzeichentest
> – Vergleich von Varianzen, F-Verteilung, F-Test
> – Chi-Quadrat Verteilung, Vergleich einer empirischen mit einer theoretischen Häufigkeitsverteilung, Chiquadrattest, Kontingenztafeln, Mc-Nemar Test
> – Einfache Regression, lineare Regression, Korrelation und spezielle Korrelationstechniken
> – Einfaktorielle und mehrfaktorielle Varianzanalyse, Varianzanalyse mit wiederholter Messung, Einzelvergleiche, Voraussetzungen der einfaktoriellen VA
> – Nicht-parametrische Verfahren: Kruskal-Wallis-Test, Friedman-Test
> – Partialkorrelation, Multiple Korrelation und Regression
> – Zweifaktorielle Varianzanalyse
> – Interaktion, feste und zufällige Effekte, Einzelvergleiche
> – drei- und mehrfaktorielle Varianzanalyse
> – Varianzanalyse mit n=1
> – Mehrfaktorielle Varianzanalyse mit wiederholter Messung
> – Multikollinearität
> – Schrittweise Regression
>
> £

Wir arbeiten mit zwei Subsets. Ein Subset mit 27 Items kam schon in Kapitel 18 zum Einsatz, das andere mit 41 Items wurde randomisiert aus der Datenbank gezogen. Die TargetWords werden bestimmt, ohne Rücksicht auf die Textart. Mit den Subsets werden nun die Karten gerechnet: einmal ohne ÜK (d.h. die TargetWords in den Listen werden wie gewöhnlich mit der 5er Hofgrösse behoft) und vier Mal mit einem gewichteten ÜK von 0.5, 1.0, 2.0 und 10.0.

19.4 Resultate

19.4.1 Subset 1

Abbildung 80 zeigt die semantische Karte, die entsteht, wenn die Hofmethode ohne Rücksicht auf die Textart die TargetWords behoft. Es sind in der Einfärbung drei Cluster zu erkennen, die den drei

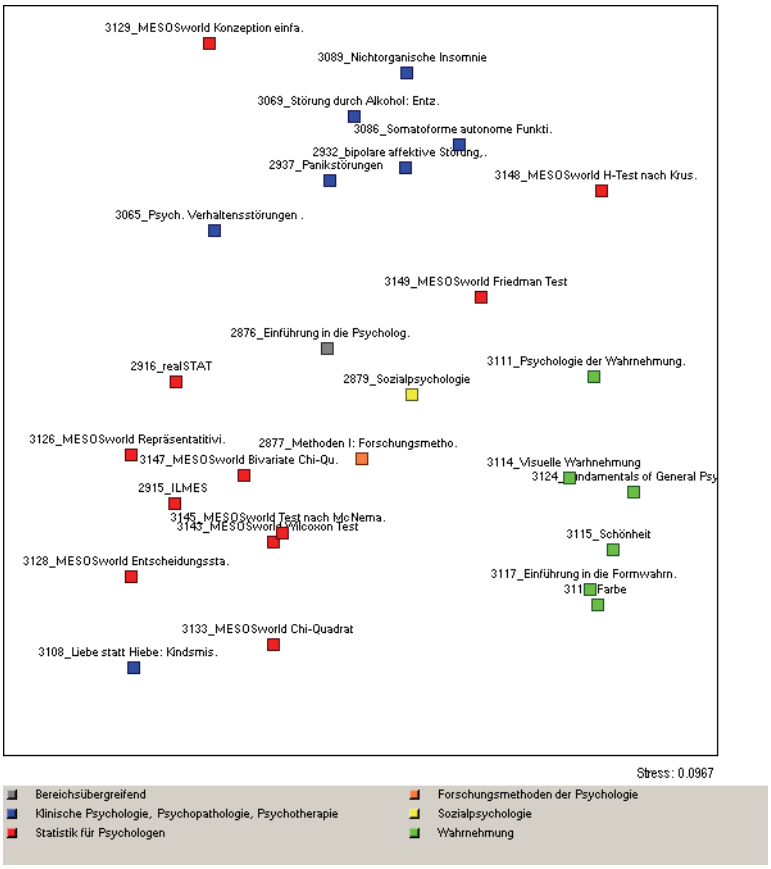


Abbildung 80: Subset 1, ohne ÜK

Kategorien «Klinische Psychologie» (blau), «Statistik» (rot) und «Wahrnehmung» (grün) entsprechen. Im Statistik-Cluster befindet sich das klinische (und exotische) Item 3108 «Liebe statt Hiebe» und im Klinik-Cluster das Statistik-Item 3129, was eindeutig falsch ist. Im mittleren Bereich sind zwei Kategorie-fremde Items und zwischen Wahrnehmung und Klinik nochmals zwei Statistik-Texte. Insgesamt ist die Karte gut strukturiert, jedoch müssten die drei erwähnten Statistik-Texte näher bei ihrem Cluster sein.

In Abbildung 81 ist dasselbe Subset dargestellt, jedoch mit unterschiedlichen Gewichtungen des ÜK. Die Auswirkungen sind minim. Erst bei einem Gewicht von 10 springt das Item 3129, jedoch landet es im falschen Cluster. Die Gesamtstruktur wirkt bei den Gewichten 0.5, 1 und 2 etwas «kerniger»; die Kategorien sind deutlicher abgegrenzt, als ohne ÜK, beziehungsweise ÜK-Gewicht 10. Der Stress steigt mit zunehmenden ÜK-Gewicht.

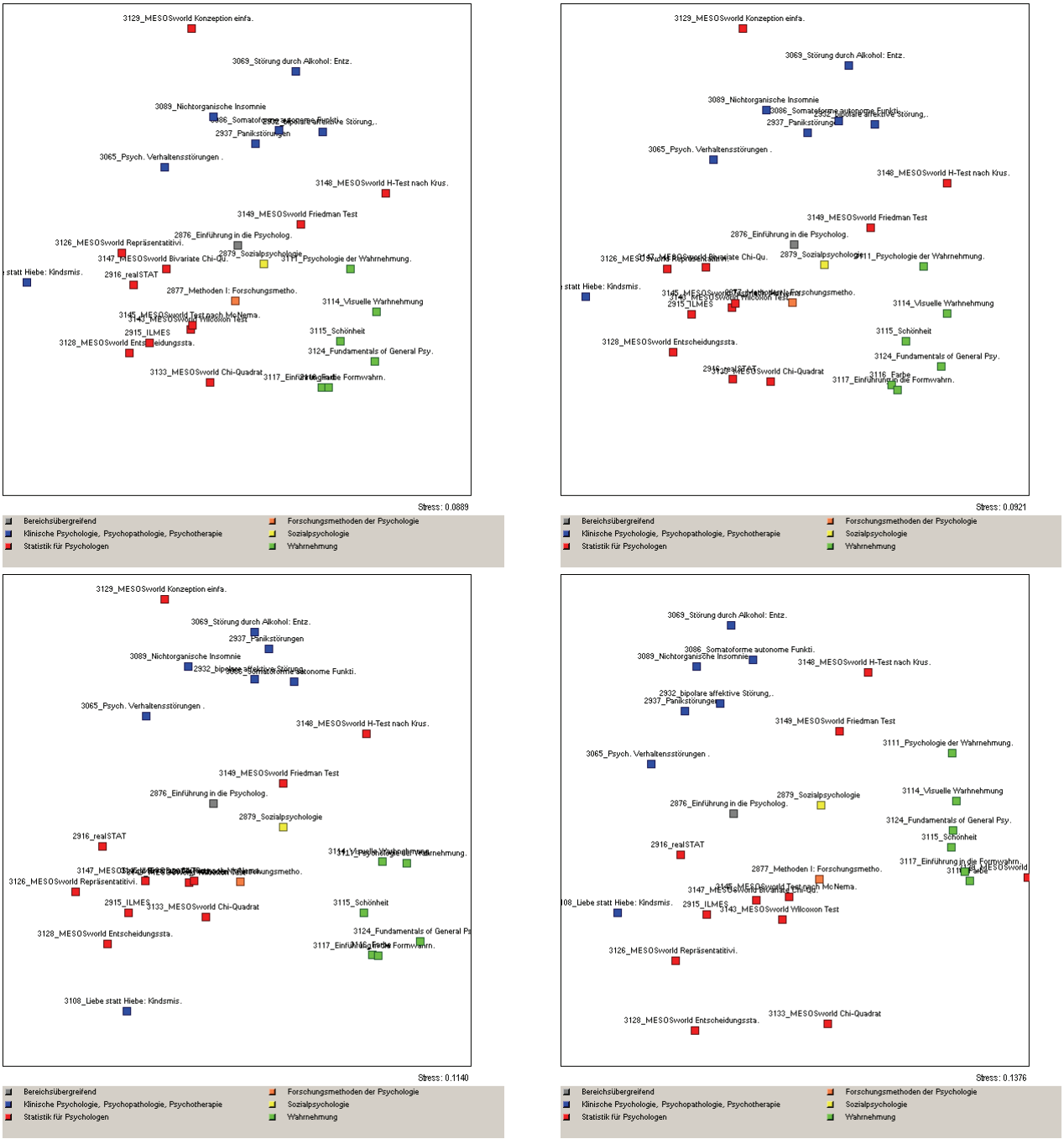


Abbildung 81: Oben links: Subset 1, ÜK-Gewicht 0.5; oben rechts: ÜK-Gewicht 1.0; unten links: ÜK-Gewicht 2.0; unten rechts: ÜK-Gewicht 10.0

19.4.2 Subset 2

Subset 2 besteht aus ziemlich heterogenen Texten. Dementsprechend fällt die Karte aus (s. Abb. 82). Die Struktur wird von den klinischen Texten dominiert (linke Kartenhälfte). In der rechten Kartenhälfte sind die Statistik-Texte, dazwischen ein Konglomerat von den restlichen Texten. Die Struktur stimmt recht

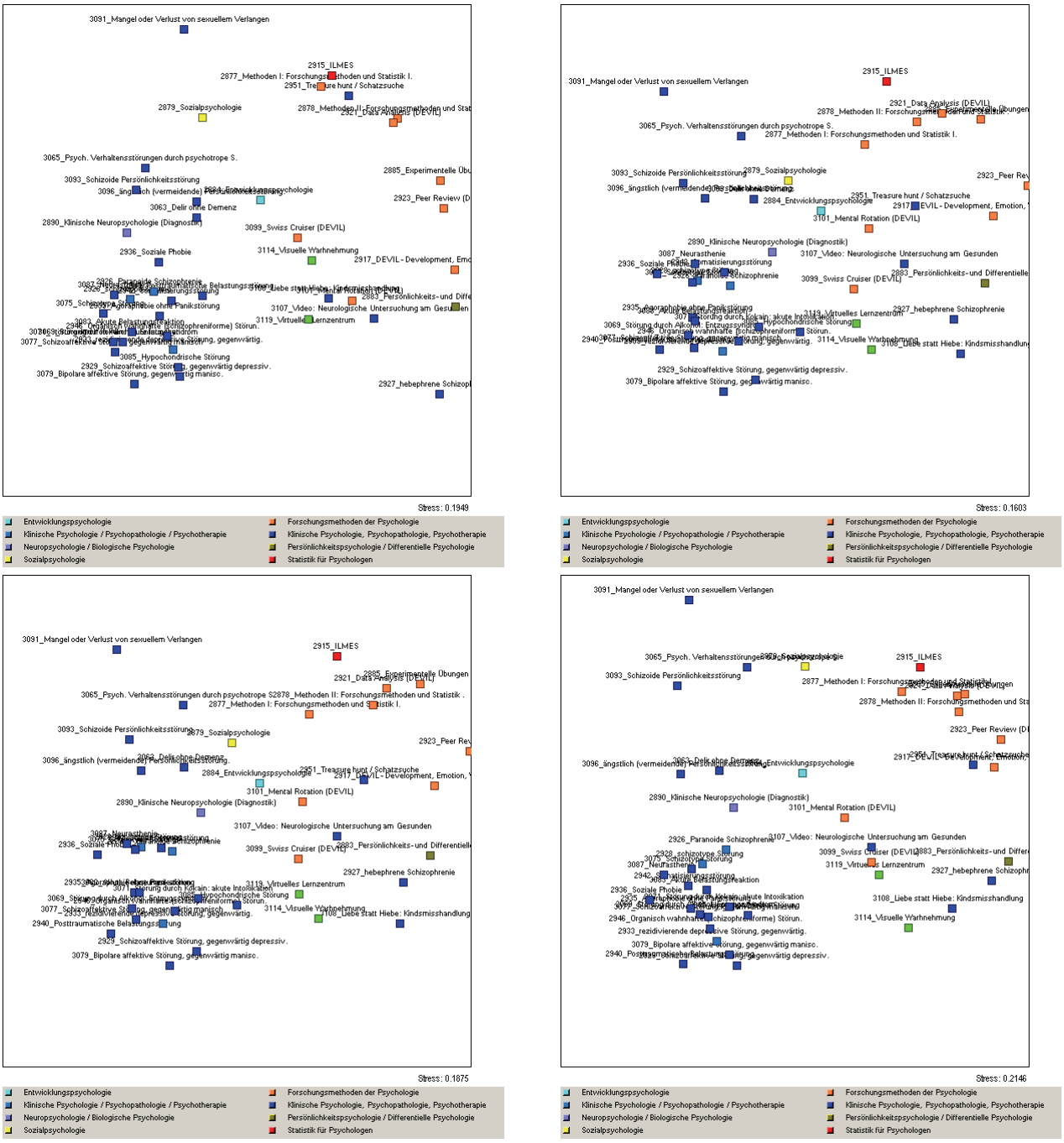


Abbildung 82: Oben links: Subset 2, ohne ÜK; oben rechts: ÜK-Gewicht 0.5; unten links: ÜK-Gewicht 1.0; unten rechts: ÜK-Gewicht 2.0

gut, jedoch sind zwei klinische Items offensichtlich falsch platziert: Item 3091, «Mangel von sexuellem Verlangen» und Item 2927 «Hepephrene Schizophrenie».

Unter Berücksichtigung des ÜK rutscht Item 3091 näher zum Klinik-Cluster und Item 2927 bildet mit drei anderen klinischen Texten einen eigenen Bereich. Die Karte wird somit etwas übersichtlicher. Bei einem ÜK-Gewicht von 10 clustern die Bereiche am stärksten, jedoch steigt auch der Stress.

19.5 Diskussion

Die gesonderte Behandlung von Listen wirkt sich in den semantischen Karten nur minim aus. Der Stress sinkt und die Struktur wird etwas prägnanter. Nach den Erkenntnissen dieses Kapitels wird ein ÜK-Gewicht von 1.0 empfohlen: Der Anteil der Liste am Gesamttext wird als solcher nicht weiter gewichtet.

20 Kurzbericht «Französische Texte»

20.1 Einleitung

Die Hofmethode wurde vor allem im deutschsprachigen Umfeld getestet. Zwar konnte das Funktionieren der HM auch mit englischem Sprachmaterial gezeigt werden, siehe Michel, 2007 (s. auch Anhang 2), jedoch blieb es bei dieser einen Fremdsprache. Theoretisch müsste die HM mit allen lateinischen Sprachen korrekt funktionieren, so auch mit Französisch, welches wir in diesem Kurzbericht testen.

20.2 Vorgehen

In der edulap-Datenbank (s. Kap. 23, Projekt edulap) befinden sich 138 französische Texte, weshalb wir gleiche diese Datenbasis verwendeten. Diese Texte wurden entrauscht, wobei eine französische Stoppwort-Liste aus dem Snowball-Projekt (snowball.tartarus.org) zum Einsatz kam. Das von uns selbst entwickelte Entrauschprogramm wurde in wenigen Details modifiziert, um auf sprachtypische Eigenheiten Rücksicht zu nehmen (z.B. Löschen von «d'»). Hierbei ist anzumerken, dass die *accents* beibehalten wurden, wie schon im Deutschen die Umlaute.

Die Keywords wurden mit Hilfe des KeywordII-Verfahrens (s. Kap. 15, KeywordII-Analyse) extrahiert, danach wurden alle Texte behoft, die Dreiecksmatrix erstellt und die NMDS gerechnet.

Um die Qualität der Kartenstruktur zu beurteilen, wurden alle Items nach der psychologischen Kategorie eingefärbt, die sie in den Metadaten zugewiesen bekommen hatten. Bei den französischen Items war diese Information jedoch nur bei rund der Hälfte der Fälle vorhanden, die andere Hälfte erscheint in den Karten weiss. Zudem entstammten die Kategoriebezeichnungen unterschiedlichen Entwicklungsstadien der Datenbank; für eine Kategorie gab es teilweise mehrere Bezeichnungen (z.B. «Forschungsmethoden» und « Forschungsmethoden der Psychologie»). Dem wurde Rechnung getragen, indem ähnlichen Kategorien die gleiche Farbe zugewiesen wurde.

20.3 Resultate

Wird der Koeffizient «Keywords pro Text» auf den bisher üblichen Wert von 0.08 gesetzt, wird die Karte nur schlecht geclustert. Mit dem relativ hohen Wert von 0.14 hingegen bilden sich deutliche Konzentrationen von psychologischen Kategorien (s. Abb. 83).

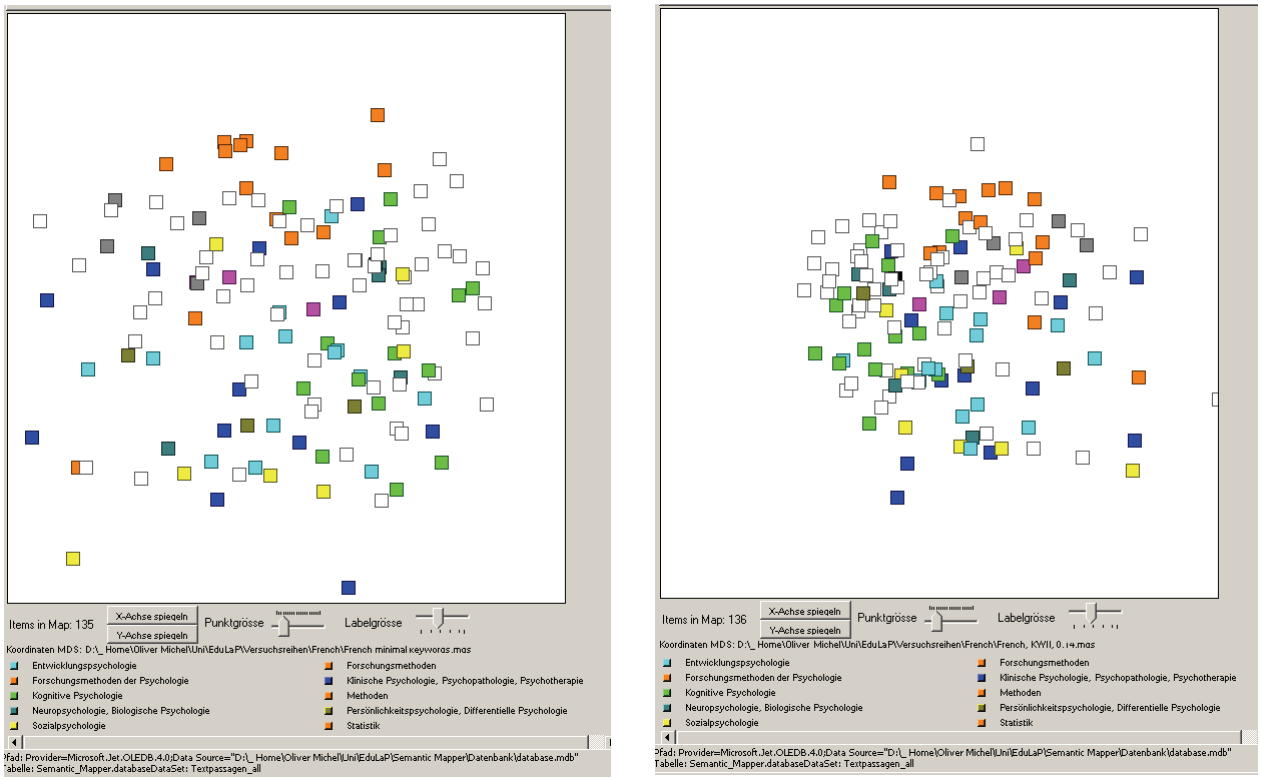


Abbildung 83: Links wurde für die KeywordII-Berechnung einen Koeffizienten von 0.08 gewählt, rechts wurde er auf 0.14 erhöht. Diese Erhöhung wirkt sich positiv auf die Clusterung aus, die enger wird.

Bei einem Koeffizienten von 0.14 werden 22 Keywords gefunden (s. Tabelle 11), bei 0.08 14 Keywords.

Tabelle 11: Bei einem KeywordII-Koeffizienten von 0.14 werden bei 138 Items 22 Keywords gefunden.

Keywords
développement
psychologie
seront
modèles
ainsi
sera
méthodes
processus
différents
objectif
recherche
évaluation
différentes
concepts
théories
chez
étudiants
introduction
présentation
bases
base
différences

20.4 Diskussion

Die Hofmethode funktioniert auch für die französische Sprache wie gewohnt. Ob der erhöhte KeywordII-Koeffizient ein systematischer Unterschied zu Deutsch oder Englisch ausdrückt, kann mit den vorliegenden Daten nicht gesagt werden.

21 Multilinguality

21.1 Überblick

Multilinguality nennen wir unser Verfahren, verschiedensprachliche Texte in einer gemeinsamen semantischen Karte zu integrieren. Durch eine automatische Übersetzung in verschiedene Fremdsprachen und einer darauf folgenden Mittelung der Ähnlichkeitswerte werden Ausreisser ausnivelliert. Die Semantik von Texten unterschiedlicher Sprachen kann somit erkannt und innerhalb derselben Karten angezeigt werden.

21.2 Einleitung

In diesem Kapitel beschreiben wir ein Verfahren, das die Flexibilität im Umgang mit heterogenen Datenquellen ungemein erhöht: Multilinguality/Mehrsprachigkeit. Unter diesem Stichwort verstehen wir die automatisierte Übersetzung von Texten in mehrere Fremdsprachen.

In der Praxis zeigt es sich immer wieder, dass mehrsprachige Texte in einem Textkorpus vorkommen. Der gängige Weg ist, dass sich der Benutzer für eine Sprache entscheiden muss und er Texte nur innerhalb dieser gewählten Sprache suchen kann. Selbstverständlich dürfen die Suchbegriffe nur in derselben Sprache geschrieben sein. Möchte man aber Texte verschiedener Sprachen innerhalb einer Karte darstellen, müssen semantische Relationen über Sprachgrenzen hinweg berechnet werden können.

Wir realisieren das, indem wir die Texte maschinell übersetzen. Verschiedene Webdienste[24] lassen sich über ein API ansprechen und übersetzen Texte in eine gewünschte Zielsprache. Was für den Alltag oft unbrauchbar ist und noch immer häufig Schmunzeln oder Sorgenfalten hervorruft, lässt sich für die HM hervorragend nutzen. Die HM ist weniger auf syntaktische Korrektheit angewiesen, als der menschliche Leser. Es genügt, wenn relevante Wörter im Hof vorkommen. Diese kommen meist auch in Übersetzungen vor, die der menschliche Leser als falsch oder zumindest stilistisch unschön abwerten würde.

24 Wir benutzten während der Entwicklungsphase des SemanticMappers GoogleTranslate (http://code.google.com/intl/sv/apis/language/translate/overview.html), das mittlerweile jedoch kostenpflichtig ist. Als Alternative zogen wir den Übersetzungsservice von Microsoft in Betracht (http://www.microsofttranslator.com/dev/).

Ein Fehler wäre es, sämtliche Texte in nur eine Zielsprache – z.B. Englisch – übersetzen zu lassen. Es besteht die Gefahr, dass sich das übersetzte Englisch von demjenigen der original englischen Texte systematisch unterscheidet und man sich somit einen unerwünschten Bias einhandelt. Wir liessen die Texte deshalb in mehrere Zielsprachen übersetzen. Wir wählten fünf Sprachen, um einerseits die gebräuchlichen Sprachen in unserem Arbeitsumfeld abzudecken und um andererseits genügend Redundanz für die spätere Verrechnung zu erhalten. Deutsche Texte werden nach Englisch, Französisch, Italienisch und Spanisch übersetzt, vice versa Texte der anderen Sprachen in die jeweils vier Alternativsprachen. Weitere Fremdsprachen kamen in unseren Untersuchungen nicht vor.

Die Originalsprache wird vom Webdienst übrigens automatisch erkannt und muss nicht manuell festgelegt werden. Die deutsche Keyword-Liste übersetzten wir ebenfalls maschinell in die vier Alternativsprachen.

Möchte sich nun ein Benutzer eine Auswahl der Texte in einer semantischen Karte anzeigen lassen, wird im Hintergrund für jede der fünf Sprachen eine eigene Dreiecksmatrix gerechnet. Diese wird anschliessend gemittelt. Erst mit der gemittelten Karte wird die NMDS gerechnet. Man schlägt dadurch – so die Idee – zwei Fliegen mit einer Klappe:

1. verschiedensprachige Texte können innerhalb derselben Karten angezeigt werden
2. Robustheit durch Mittelung

Es geschieht manchmal, dass ein Item in der semantischen Karte falsch platziert wird. Die Gründe können vielfältig sein, beispielsweise ein zu kurzer Text oder zu wenige TargetWords. Manchmal ist der Grund aber ein irreführendes, z.B. zweideutiges, Wort im Kontext. Da es unwahrscheinlich ist, dass sich ein Übersetzungsfehler dieser Art in anderen Sprachen wiederholt, kann er durch die Mittelung ausgemerzt werden. Die Karte wird dadurch robuster.

Wir haben das Verfahren im edulap-Projekt (s. Kap. 23, Projekt edulap) implementiert. Die Datenbank bestand zum Zeitpunkt der Implementation aus 325 Items, von denen 186 deutsch waren, 138 französisch und einer englisch. Wir untersuchten zwei Aspekte der Multilinguality:

21.2.1 Qualitätsvergleich der übersetzten Karten

Wie wirkt sich die automatisierte Übersetzung auf die Qualität der semantischen Karte aus: Entstehen in einzelnen Sprachen Artefakte oder werden bestehende Fehler ausgemerzt? Gewinnt die gemittelte Karte an Semantik oder verliert sie?

21.2.2 Strukturvergleich: Kombination zweier Sprach-Konfigurationen

Was passiert, wenn zwei eigenständige Kartenkonfigurationen unterschiedlicher Sprachen zusammengelegt werden? Werden die beiden Sprachen gut durchmischt und die Texte nach ihrer Semantik platziert oder brechen die Konfigurationen zusammen und entsteht so eine unerwünschte Karte der Sprachen?

21.3 Vorgehen

21.3.1 Qualitätsvergleich

Ein Subset von je drei Items aus acht verschiedenen psychologischen Richtungen wurde zusammengestellt. Alle Texte waren original deutschsprachig und wurden in die Sprachen Englisch, Französisch, Italienisch und Spanisch übersetzt. Für jede einzelne Sprache, sowie für die gemittelten Ähnlichkeitswerte, wurde eine Karte erstellt. Drei Mitarbeitern des psychologischen Instituts wurden die sechs Karten präsentiert, wobei die Texte mit ihren Titeln angeschrieben waren, jedoch nicht mit ihrer psychologischen Fachrichtung. Die MA wurden gebeten, die Karten in eine Rangfolge zu bringen, die das Mass der Sinnhaftigkeit wiedergab.

Ziel war herauszufinden, ob eine bestimmte Karte immer am Anfang oder Ende der Rangliste platziert wird. Beispielsweise könnte es sein, so die Hypothese, dass die gemittelte Karte immer am schlechtesten abschneidet und die originalsprachige am besten. Die einzelnen Karten wurden untereinander prokrustet, um die Unterschiedlichkeit beurteilen zu können. Im Sinne einer explorativen Studie wurde das Vorgehen nur mit einem Itemset gemacht. Alle Texte waren separat ersichtlich, falls ein Titel nicht aussagekräftig genug war und die MA den Inhalt nachschlagen wollten.

21.3.2 Strukturvergleich

Eine Karte mit 24 original deutschsprachigen Items aus acht Kategorien, sowie eine zweite Karte mit 24 original französischsprachigen Items aus denselben acht Kategorien wurden erstellt. Es ist zu erwarten, dass beide Karten für sich eine sinnvolle Struktur bilden. Dann wurden die 48 Items automatisiert in die anderen Sprachen (DE/FR, EN, IT, SP) übersetzt, Ähnlichkeiten gerechnet, gemittelt und die gemittelte Karte erstellt.

21.4 Resultate

21.4.1 Qualitätsvergleich

In der deutschen Karte fallen zwei Items auf, die offensichtlich falsch platziert sind: «Podcast Einführung in die Sozialpsychologie» und «Multiple Persönlichkeitsstörung» (s. Abb. 84). Beim Podcast dürfte der Grund darin liegen, dass das Abstract den Inhalt nur kurz beschreibt. Der grössere Teil beinhaltet Informationen über die Strukturierung des Kurses:

> Die Vorlesung soll in die zentralen Themenbereiche der Sozialpsychologie einführen, aber auch die Gelegenheit geben, anhand von Filmen das eine oder andere klassische Experiment der Sozialpsychologie kennen zu lernen und inhaltliche Fragen zu stellen. Die Vorlesung ist obligatorisch für das propädeutische Jahr im Rahmen des Bachelorstudiums. Grundlagentexte der Vorlesung sind einzelne Kapitel aus vier verschiedenen Lehrbüchern, die als pdf-Dateien zugänglich gemacht werden. Es ist das Ziel, diese Literatur didaktisch aufzubereiten und möglichst anschaulich zu vermitteln. Ergänzend zu dieser Einführungsvorlesung wird im HS 2009 eine weitere Veranstaltung (Vorlesung/Arbeitsgruppe: Sozialpsychologie II) stattfinden, in der ausführlich auf die verschiedenen Anwendungsbereiche der Sozialpsychologie eingegangen wird, wie z.B. Gesundheits-, Umwelt- oder Rechtspsychologie.

Die Missplatzierung des Items «Multiple Persönlichkeitsstörung» ist unklar (und wird im Rahmen dieses Berichtes nicht weiter verfolgt).

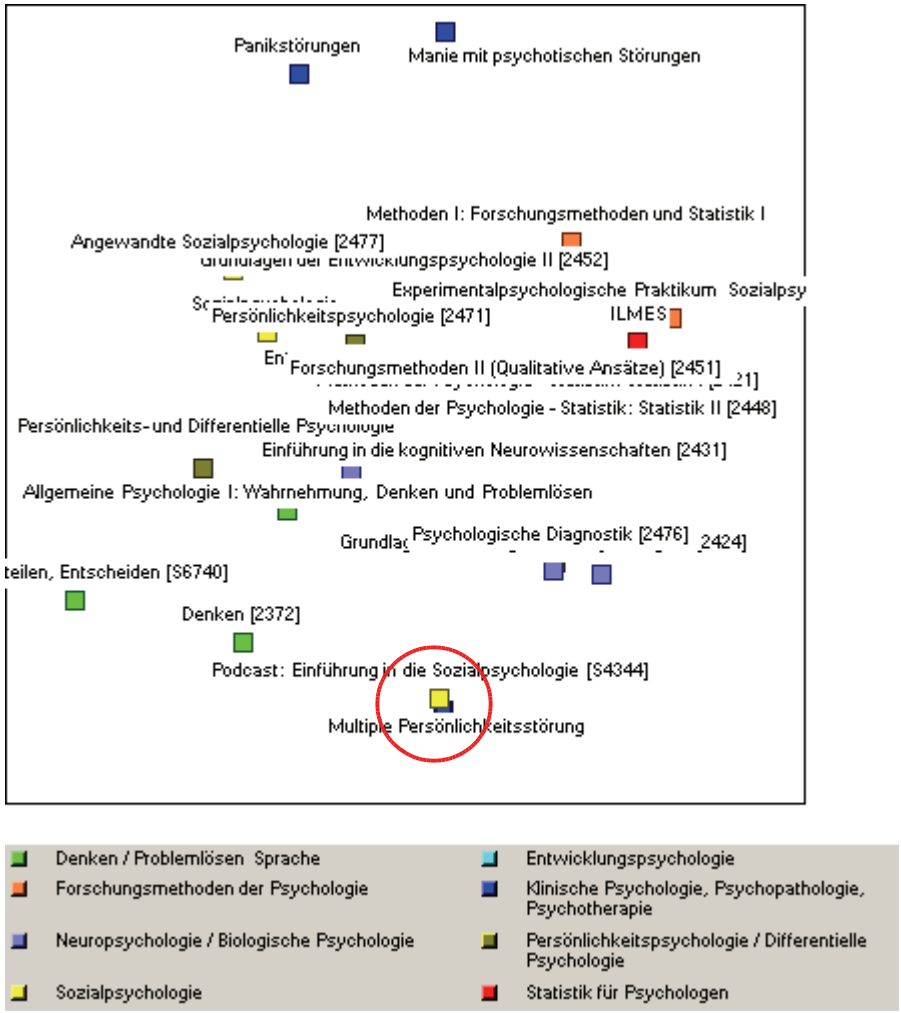


Abbildung 84: Die beiden eingezeichneten Items sind deutlich falsch platziert.

Die Karten der übersetzten Sprachen (s. Abb. 85) zeigen, dass es in jeder Sprache einzelne Ausreisser gibt (ausser bei EN und IT). Erst bei der gemittelten Karte sind die Ausreisser ausnivelliert, trotzdem sind keine deutlichen Cluster erkennbar (s. Abb. 86).

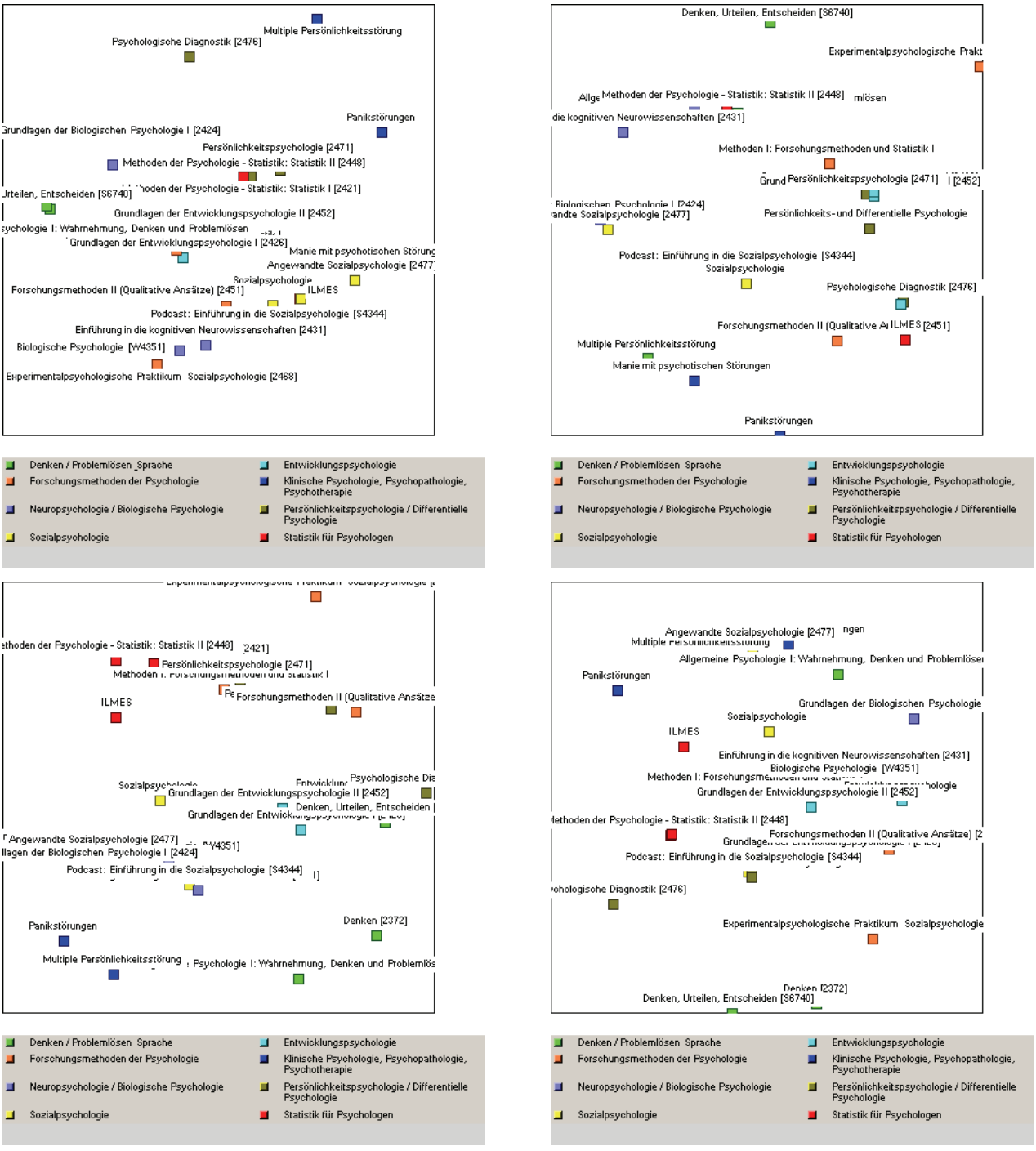


Abbildung 85: Die Karten der einzelnen Sprachen. Im Uhrzeigersinn: Englisch, Französisch, Italienisch, Spanisch.

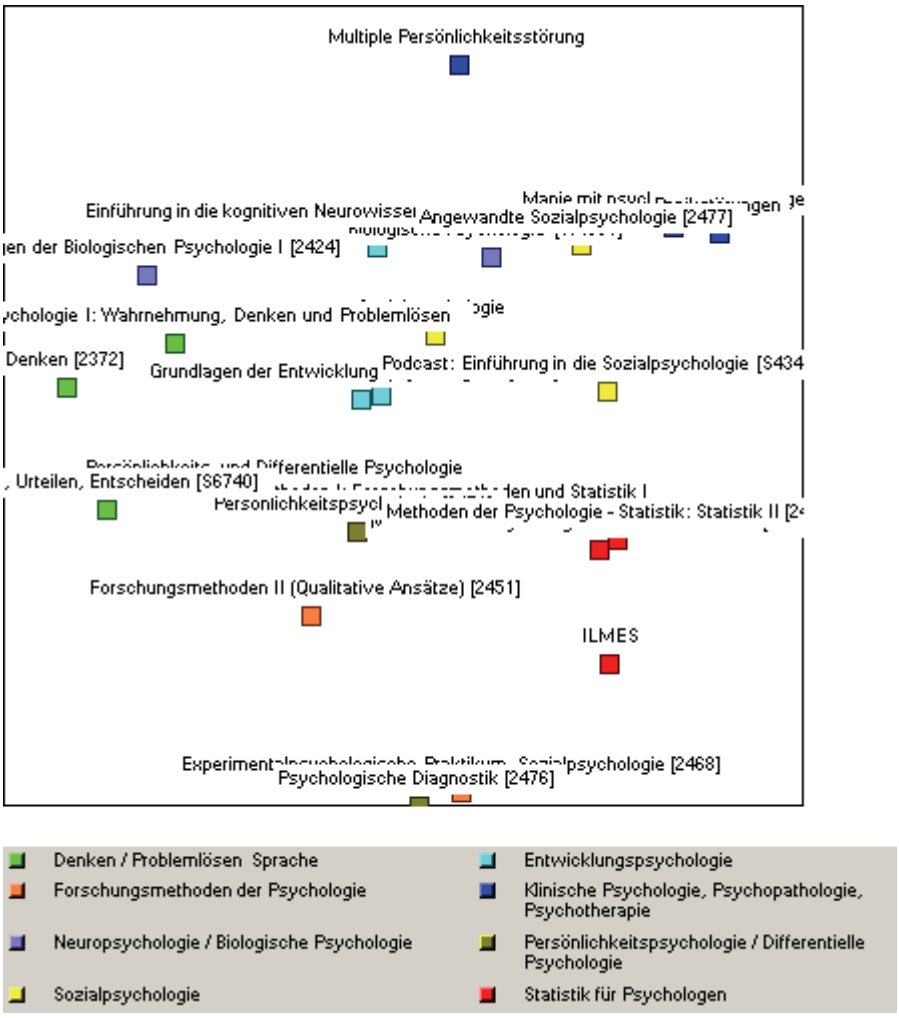


Abbildung 86: Die gemittelte Karte: Keine Ausreisser, jedoch auch keine Cluster.

Auch die Ranglisten der MA zeigen keine eindeutigen Präferenzen (s. Tab. 12). Die deutsche Originalkarte wird zwar von einer Person präferiert, ist bei den beiden anderen aber erst an vierter Stelle. Die gemittelte Karte ist bei einer Person an letzter, bei den beiden anderen jedoch an erster, beziehungsweise zweiter Stelle. Mittelt man die Rangplätze, ist die deutsche Karte zusammen mit der gemittelten an erster Stelle, jedoch ist das bei drei VP nur bedingt aussagekräftig. Dennoch erstaunt, wie heterogen die Karten bewertet werden. Einschränkend muss gesagt werden, dass alle Karten durch die langen Titel recht unübersichtlich waren (s. Abb. 87).

Tabelle 12: Die Rangordnungen von drei VPn ergeben kein homogenes Bild: Keine Karte wird deutlich präferiert oder verworfen. Der gemittelte Rang ist nur bedingt aussagekräftig.

	VP 1	VP 2	VP 3	gemittelter Rang
1. Rang	DE, EN	FR	gemittelt	DE, gemittelt
2. Rang		gemittelt	SP	EN, FR
3. Rang		EN	IT	SP
4. Rang	FR, SP, IT	DE	DE	IT
5. Rang		SP	FR	
6. Rang	gemittelt	IT	EN	

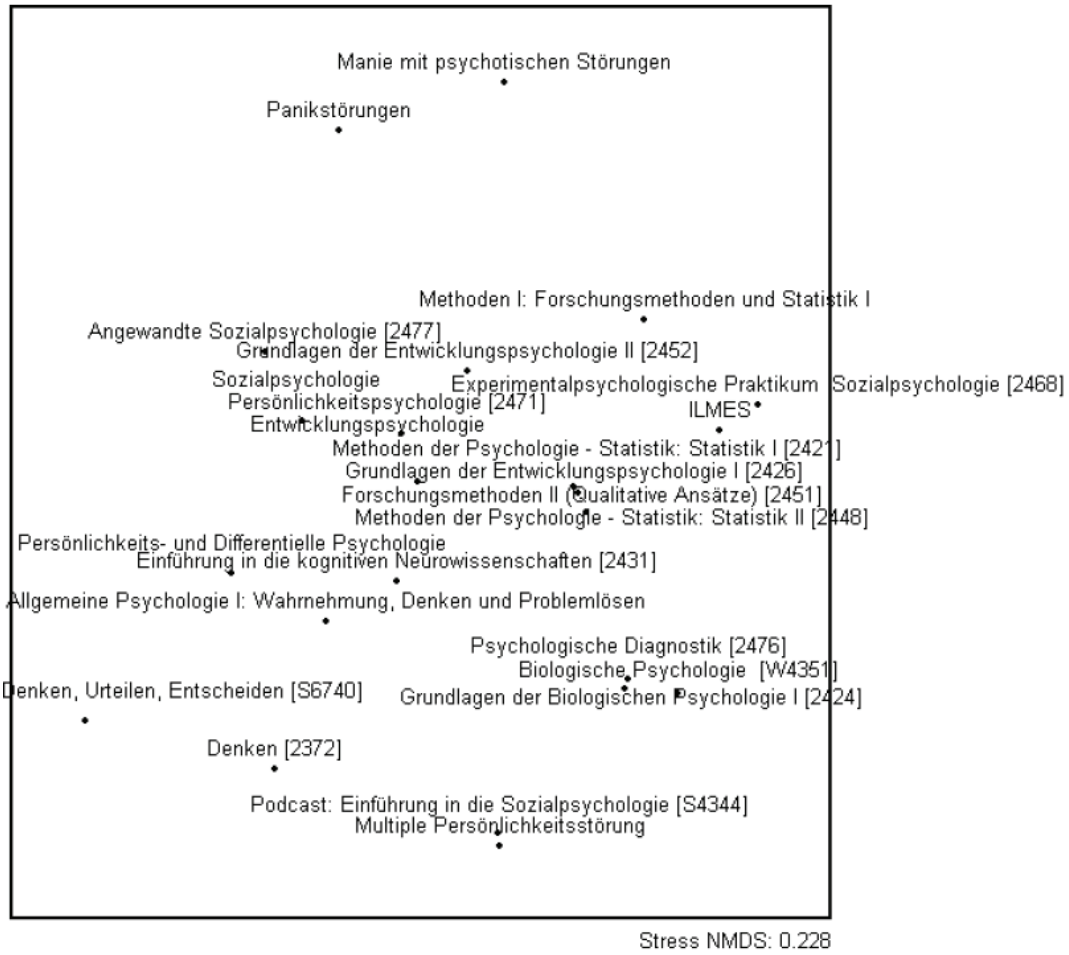


Abbildung 87: Beispiel einer Karte, wie sie den VP gezeigt wurde (Massstab 1:1)

Die prokrusteten Karten zeigen, dass sich die Konfigurationen zwischen den Sprachen deutlich unterscheiden (s. Anhang 4).

21.4.2 Strukturvergleich

Die beiden Strukturen (DE und FR) sind insgesamt recht unterschiedlich (s. Abb. 88): Bei beiden formiert sich zwar ein Denken- und ein Statistik/Methoden-Bereich und bei beiden sind die Persönlichkeitspsychologie-Items verstreut, in folgenden Bereichen unterscheiden sie sich aber: Bei der deutschsprachigen Karte ist der Entwicklungsbereich in der Kartenmitte, zwischen Denken und Statistik, bei der französischen aber schiebt sich Denken zwischen Entwicklung und Statistik. In der französischen Karte ist zudem die Neuropsychologie und die Klinische Psychologie über den gesamten Raum verteilt.

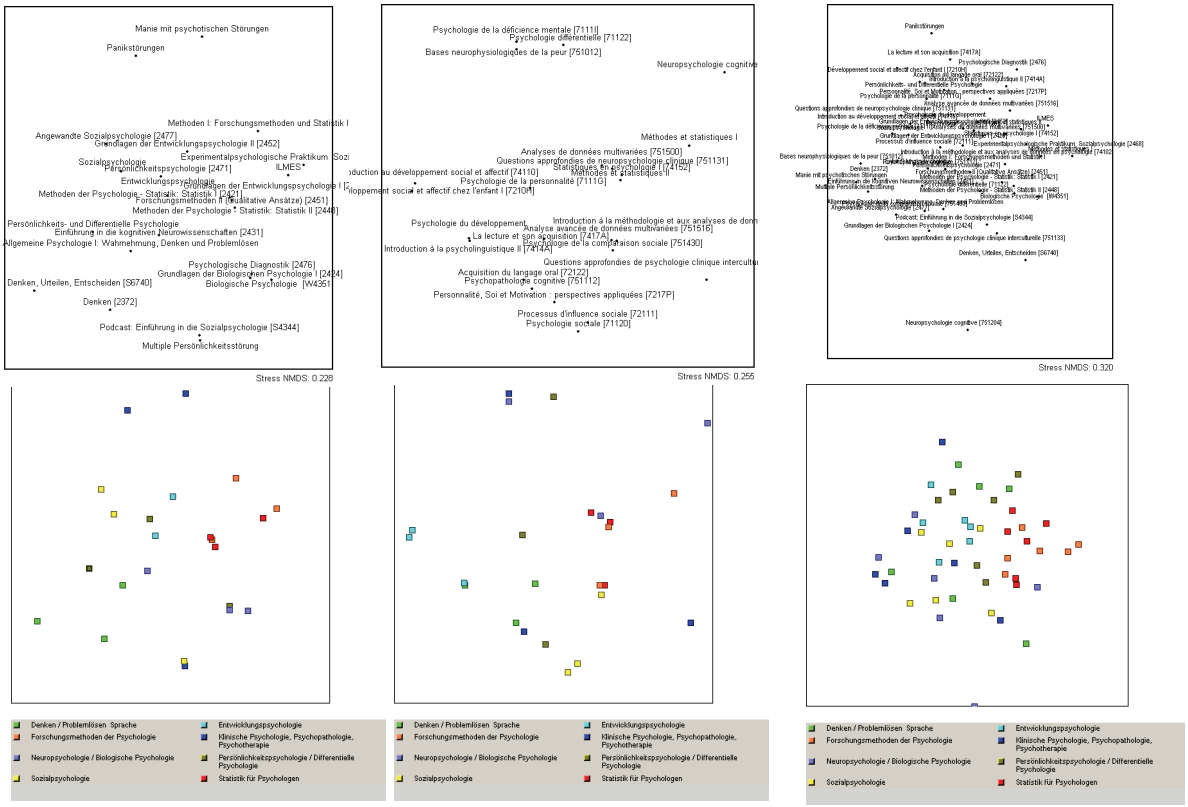


Abbildung 88: Links ist die Karte mit den deutschsprachigen Items, in der Mitte die französischsprachigen, rechts die mit den beiden Sprachen integrierte und gemittelte Karte. In der oberen Reihe sind die Items mit den Titeln angeschrieben, in der unteren Reihe sind dieselben Karten, jedoch wurden die Items nach Kategorie eingefärbt.

In der zusammengelegten Karten zeigt sich nochmals ein anderes, enttäuschendes Bild: Einzig der Statistikbereich bleibt beisammen, die Sozialpsychologieitems und die Denkitems wurden aufgespalten, Neuro-, Klinische und Persönlichkeitspsychologie sind ziemlich weit verteilt.

Positiv hervorzuheben ist, dass die unterschiedlichsprachigen Items in der integrierten Karte gut durchmischt werden. Die Items werden tatsächlich innerhalb der Themenbereiche über die Sprachgrenzen hinweg semantisch angeordnet. Einzig das zweigeteilte Denken-Cluster ist auch sprachen-geteilt. Trotzdem ist eine gewisse Nord-Südverteilung auf der Karte (s. Abb. 89) auszumachen. Worauf diese zurückzuführen ist, kann im Rahmen dieses Experimentes nicht gesagt werden, verspricht aber weitere, spannende Untersuchungen.

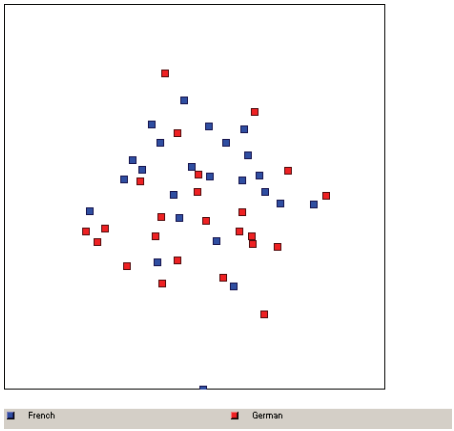


Abbildung 89: Sprachvergleich: gute Durchmischung innerhalb der Cluster, aber eine Nord-Südverteilung?

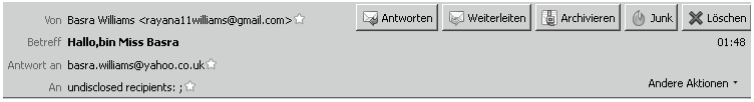
21.5 Diskussion

Die automatische Übersetzung bringt tatsächlich eine gewisse Stabilisierung der Karten mit sich. Ausreisser werden ausnivelliert, ohne dass es aber zu einer stärkeren Clusterung kommt. Die Konfigurationen der einzelnen Sprachen sind stark unterschiedlich. Die deutsche Originalkarte sowie die gemittelte Karte scheinen den Fremdsprachen gegenüber leicht präferiert zu werden, jedoch ist die Datengrundlage mit diesem einen Itemset und den drei VP zu gering, um systematische Unterschiede der Sprachen aufzuzeigen.

Der Strukturvergleich der deutschen und französischen Karten zeigt ein interessantes Bild: Zwar sind die getrennten Strukturen sehr unterschiedlich und noch mehr diejenige der gemeinsamen Karte, dennoch mischen sich die Texte innerhalb eines Themenbereichs sehr gut.

Die Multilinguality ist ein mächtiges Feature, das in der Praxis viele Vorteile bringen wird. Aus wissenschaftlicher Sicht bleibt aber noch viel zu erforschen.

Teil V Experimente und Anwendungen



Hello,Am miss Basra,interested in you,and wish to have you as my friend,for a friend is all about Respect,Admiration and love passion also friendship is consist of sharing of ideas and planing together,I have a special something I want to discuss with you,I do not care were my kind of friend comes from, provided he/she is a nice and caring person,Basra.

.....

Hallo,bin Miss Basra,an Ihnen interessiert,und möchte dich als meinen Freund zu haben,für einen Freund dreht sich alles um Respekt,Bewunderung und Liebe Leidenschaft auch Freundschaft bestehen aus Teilen von Ideen und Hobeln zusammen ist,habe ich eine besondere etwas,was ich will mit Ihnen diskutieren,ist mir egal,waren meine Art von Freund kommt aus,vorausgesetzt,er/sie ist eine nette und fürsorgliche Person,Basra.

Spam-Mail, erhalten am 31.07.12

22 Bedeutungsähnlichkeiten von Abstracts: Vier Verarbeitungsebenen

22.1 Grundidee

Am 13./14. September 2007 fand in Zürich der SGP-Kongress mit über 300 Referenten, beziehungsweise Beiträgen statt. Anlässlich dieser Veranstaltung untersuchten wir vier unterschiedliche Verarbeitungsebenen von automatisierten Methoden zur Textähnlichkeitsberechnung der eingereichten Abstracts.

Im Zusammenspiel mit einer intuitiv zu verstehenden Visualisierung mittels einer semantischen Karte ergeben sich folgende Vorteile einer erfolgreichen automatisierten semantischen Texterkennung:

(i) Kongressorganisatoren können die eingereichten Beiträge leichter zu inhaltlich passenden Sessionen formen
(ii) die Besucher können sich anhand der Karte rasch einen Überblick der dargebotenen Beiträge verschaffen
(iii) die Teilnehmer erkennen rasch, wer in einem inhaltlich ähnlichen Gebiet tätig ist

Erkennung von Textähnlichkeiten ist eine klassische Aufgabe des Information Retrieval. Ein häufiges Problem bei der maschinellen Verarbeitung von Text ist, dass sich die Textanalyse am Wortbild festmacht: Was gleich heisst, ist gleich, und vice versa: Was anders benannt wird, ist nicht gleich. Anhand der damit verbundenen Synonym-/Homonymproblematik zeigen sich grundlegende Limitierungen der oberflächlichen Textanalyse: Die Bedeutung von Wörtern zeigt sich eben nicht ausschliesslich in ihrer Form, vielmehr wandelt sie sich mit jedem Gebrauch und lässt sich nur durch ihren jeweils spezifischen Kontext erfassen. Inwieweit sich diese Erkenntnis nutzbringend in eine algorithmische Form bringen lässt, ist Thema dieses Experiments.

22.2 Die vier Methoden

Die folgenden Methoden wurden angewendet:

(i) Trigrammierung
(ii) Überlappungskoeffizient

(iii) Stichwortmethode

(iv) Hofmethode

(i) Trigrammierung

Bei der Trigrammierung wird der Text in 3er-Gruppen von Buchstaben unterteilt, wobei dieses Gruppenfenster jeweils um einen Buchstaben verschoben wird: Das erste Trigramm besteht aus dem ersten, zweiten und dritten Buchstaben des Textes, das zweite Trigramm aus dem zweiten, dritten und vierten Buchstaben, etc. Anschliessend wird zwischen zwei Texten auf die Ähnlichkeit geschlossen, indem der Anteil gemeinsamer Trigramme berechnet wird.

Diese Methode ist somit vollständig losgelöst von der syntaktischen Situation und der semantischen Struktur.

(ii) Überlappungskoeffizient

Der Überlappungskoeffizient (ÜK) (Marx, 1976) berechnet den Anteil gemeinsamer Wörter zweier Texte (relativ zur jeweiligen Textlänge). Er wurde ursprünglich als Mass für die assoziative Bedeutungsähnlichkeit mehrerer Reize entwickelt.

Der ÜK behandelt die Wörter zwar als Ganzes, jedoch isoliert und aus dem Zusammenhang gerissen, ähnlich dem Vector Space Model (Salton, 1975; Panyr, 1986).

(iii) Stichwortmethode

Im Gegensatz zum ÜK werden bei der Stichwortmethode nicht sämtliche Wörter auf ein gemeinsames Vorkommen verglichen, sondern nur eine Auswahl. Diese Auswahl wurde im Zusammenhang mit der folgenden Hofmethode erarbeitet und wird im nächsten Abschnitt beschrieben.

Auch die Stichwortmethode arbeitet ohne Kontext, jedoch in dem Sinne semantisch, als dass nur mehrheitlich bedeutungsvolle Wörter in die Ähnlichkeitsberechnungen einfliessen.

(iv) Hofmethode

Die Hofmethode wurde im einleitenden Teil bereits ausführlich beschrieben.

22.3 Das Experiment

Der Vergleich dieser vier Methoden soll zeigen, wie sich abnehmender Kontexteinbezug auf die Ähnlichkeitsberechnung auswirkt: Die Hofmethode bezieht die umgebenden Wörter bestimmter Stichwörter mit ein; die Stichwortmethode betrachtet nur die Stichwörter selbst, reisst sie also aus dem Kontext; der Überlappungskoeffizient kennt keine Stichwörter, sondern gewichtet alle Wörter eines Textes gleich; die Trigrammierung arbeitet nicht mal mit Wörtern, sondern nur mit der Überlappung von unsortierten Buchstabenfolgen.

Bei allen vier Methoden besteht die Textgrundlage aus den eingereichten Abstracts. Die Ähnlichkeit der Texte wird paarweise erhoben. Die resultierende Ähnlichkeitsmatrix wird anschliessend per NMDS in eine zweidimensionale Struktur überführt, die die errechneten Ähnlichkeiten visuell wiedergibt.

Bei der Stichwort- und der Hofmethode müssen irgendwie die zu vergleichenden Stichworte bestimmt werden. Das geschieht in diesem Experiment über die Titel und Untertitel: Manuell wurden deren Nomina bestimmt und in einem Pool gesammelt. Es ergaben sich knapp 800 Nomina. Für den paarweisen Vergleich zwischen zwei Texten (Text A und Text B) wird entweder das Auftreten an sich (Stichwortmethode) oder die Höfe gemeinsamer Stichwörter (Hofmethode) verrechnet. Kommt das Stichwort X in einem Text mehrfach vor, wird das hinzugerechnet. Die Werte aller Paare werden aufsummiert; diese Summe repräsentiert die Ähnlichkeit zwischen Text A und Text B. Dabei wird ein unscharfer Wortvergleich angewendet: Es reicht, wenn sich zwei Stichworte äusserlich ähnlich sind. Das geschieht mit jedem Auftreten aller Stichwörter in jedem Text.

Die resultierenden Karten müssen anhand eines Gütekriteriums gemessen werden. Neben der Augenscheinvalidität (sind themenverwandte Beiträge nahe beisammen platziert) gibt es praktischerweise das Aussenmass der Gruppenzugehörigkeit: Die Texte werden in der NMDS-Karte nach ihrer von der Kongressleitung zugewiesenen Gruppe eingefärbt. Die Texte müssten, sofern die Ähnlichkeiten richtig berechnet wurden, farblich einheitliche Cluster bilden.

22.4 Hypothese

Der Kongress besteht aus Symposien (verschiedene Beiträge zu einem Thema), Poster und Vortragsgruppen. Es ist zu erwarten, dass die Symposiumsbeiträge deutlich geclustert werden. Weiter

sollte die Hofmethode die Beiträge thematisch zutreffender ordnen als die alternativen Berechnungsmethoden.

22.5 Resultate

Um die Übersichtlichkeit der Karte zu wahren, wählten wir die Beiträge zu 10 ausgewählten Symposien aus, wobei wir eine thematisch möglichst grosse Breite herzustellen versuchten (s. Abb. 90).

Es zeigt sich, dass die Stichwort- wie die Hofmethode sehr ähnliche Karten mit hoher semantischer Ordnung produzieren, die mit der organisatorischen Gruppenzuteilung sehr gut korrespondieren, wenn auch keine separierten Cluster auszumachen sind. Jedoch sind grösstenteils alle Beiträge zu einem Symposiumsthema im selben Bereich. So bilden die sechs grauen Beiträge (Psychotherapy and more: Aspects of clinical investigations from disorder to outcome) rechts unten einen eigenen Bereich, ebenso die vier dunkelgrünen Beiträge (Active risk management in the decision process: the role of risk-defusing operators) links unten, die sechs violetten Beiträge (Current developments and research directions in positive psychology) rechts und die vier hellgrünen Beiträge (New Trends in Developmental Cognitive Psychology) in der Mitte. Die Entwicklungspsychologie passt sehr gut in diesen zentralen Bereich, beinhaltet sie doch ein Querschnittsthema der Psychologie. Die orangen und gelben Beiträge (Learning in organizations und eLearning in Swiss Psychology - Status Quo and Future Perspectives) oben mittig trennen sich hingegen nicht auf, was inhaltlich tatsächlich nicht überraschend ist. Einzig das rosarote Thema (Challenges in Group Interactions and Performance) verteilt sich über die gesamte Karte – bei der HM-Karte mehr, als bei der Stichwortkarte.

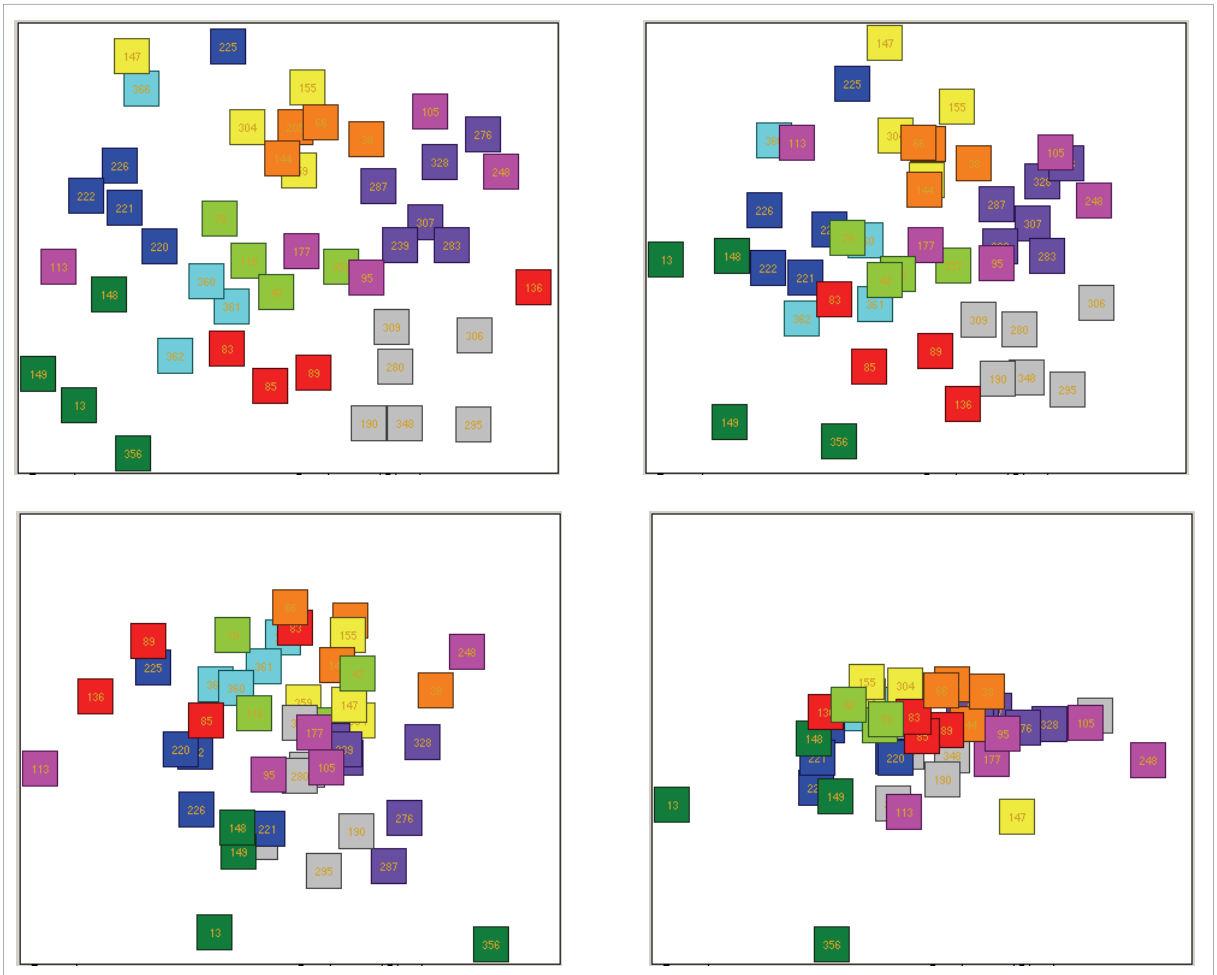


Abbildung 90: Eingezeichnet sind die Beiträge zu 10 verschiedenen Symposien, wobei die Einfärbung die einzelnen Themenbereiche wiedergibt. Links oben basiert die Karte auf Ähnlichkeiten, die mit der Hofmethode errechnet wurden, rechts oben mit der Stichwortmethode, links unten mit dem Überlappungskoeffizienten, rechts unten mit der Trigrammierung.

Die Trigrammierung, sowie der ÜK ordnen die Karten ebenfalls thematisch sinnvoll, allerdings nicht so deutlich wie die beiden erstgenannten Methoden. Vor allem die Trigramm-Karte wird in ihrer Strukturierung durch die beiden Beiträge links und unten dominiert. Da zudem beide Verfahren ziemlich rechenintensiv sind, darf man festhalten, dass sie der Hof-, beziehungsweise Stichwortmethode unterlegen sind.

22.6 Diskussion

Wir haben es hier mit einer speziellen Textgattung zu tun: Wissenschaftliche Abstracts sind extrem dichte Texte, in denen praktisch keine semantische Redundanz auftritt. Wenn nun eine genügend grosse Stichwortbasis vorhanden ist (zur Erinnerung: 800 Nomina wurden manuell erarbeitet), ist die Stichwortmethode der Hofmethode vorzuziehen, da sie effizienter arbeitet. Die Stärken der Hofmethode liegen im Umgang mit ambiguem Text und sie braucht wenige Stichworte, um genügende semantische Spezifität zu erzeugen (ein einzelnes Stichwort kann schon genügen, wie in (Michel, 2006) gezeigt).

Die Stichwortmethode ist ein Spezialfall der Hofmethode: Die Stichworte fliessen quasi einfach ohne Hof in die Ähnlichkeitsberechnung ein. Es lässt sich somit zusammenfassend sagen, dass die Hofmethode eine geeignete Methode ist, um wissenschaftliche Abstracts solcherart in semantischen Karten zu ordnen, dass Kongressorganisation wie auch Teilnehmer und Besucher in grosser Weise davon profitieren können.

22.7 Ausblick

Die Auswahl der Stichworte mittels manueller Bestimmung der Nomina in den Titeln ist zu arbeitsaufwändig, auch wenn das eine Software übernehmen könnte[25]. In den Kapiteln 14 (Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient) und 15 (KeywordII-Analyse) wurden bereits passendere Verfahren entwickelt. Eine weitere Möglichkeit wäre folgender Ansatz: Manuell werden einige Stichworte bestimmt, die domänenspezifisch sind. Die Texte werden nach diesen Stichwörtern behoft und die NMDS gerechnet. Im nächsten Schritt werden die Clusterschwerpunkte bestimmt. Die Texte innerhalb dieser Cluster liefern in ihren Höfen nun die weiteren, domänenspezifischen Stichworte, nach denen die Texte nochmals behoft werden.

25 beispw. FreeLing 2.0, http://garraf.epsevg.upc.es/freeling/

23 Projekt edulap

23.1 Überblick

Dieses Kapitel beschreibt die Implementation der Hofmethode im interuniversitären Projekt edulap, in dem digitale Lehrressourcen aus dem psychologischen Umfeld zugänglich gemacht werden. Um die Qualität der automatisiert berechneten Karten beurteilen zu können, wird eine Expertenkarte erstellt und mit dem Output der HM verglichen.

23.2 Einleitung

Die Hofmethode wurde erstmals im schweizweiten Projekt edulap (Educational Landscape Psychology) implementiert. Beim edulap-Projekt werden digitale Lehrressourcen in einer Datenbank erfasst und durchsuchbar gemacht. Die Resultate werden dabei nicht (nur) als gewohnte Liste, sondern als semantische Karte retourniert. Die Berechnung der semantischen Ähnlichkeiten erfolgt dabei durch die Hofmethode.

Ein Auszug aus der edulap-Homepage[26] bettet das Projekt in die Landschaft des «Blended Learning» ein:

> Seit einigen Jahren werden (universitäre) Lehrveranstaltungen vermehrt mit den Möglichkeiten aus dem Bereich "Neuer Medien" (eLearning) kombiniert. Das Schlagwort Blended Learning steht dabei für eine zeitgemässe didaktische Lern- und Unterrichtsform. Im Zuge dieser Entwicklung wurden zahlreiche virtuelle Lehrangebote geschaffen und erfolgreich eingesetzt. Dabei erschöpft sich eLearning nicht in grossen, vollständigen und gut bekannten Curricula, sondern auch kleine und mediendidaktisch wenig ausgefeilte Produkte können im Rahmen von Blended Learning willkommene Ergänzungen zum bisherigen Ausbildungsangebot darstellen. Letztere sind aber nur selten einer erweiterten Öffentlichkeit bekannt. Dieses steigende Angebot führt also zunehmend zu Intransparenz und Unübersichtlichkeit.
>
> Um diesen sich laufend verschärfenden Missstand zu beheben, entwickelt das Projekt edulap ein technisches Orientierungssystem mit dem Ziel, die Suche nach und das Auffinden von digitalen Lehrressourcen und dazugehöriger Lehrveranstaltungen eines bestimmten Faches auf innovative Weise zu vereinfachen. In so genannten Orientierungskarten, in welchen die inhaltliche Ähnlichkeit zwischen den Veranstaltungen und den digitalen Lehrressourcen graphisch abgebildet ist, können Dozierende und Studierende diejenigen Angebote lokalisieren, welche Ihren Bedürfnissen am besten entsprechen. Die umfangreichen von edulap zusammengetragenen Meta-Informationen zum verfügbaren Angebot im Fach Psychologie der Schweiz auf Bachelor- und Masterstufe erlauben eine schnelle und präzise Beurteilung der interessierenden Möglichkeiten. Im Speziellen liefert edulap auch Informationen über die Nutzungsmöglichkeit oder verweist direkt auf das Angebot selber (z.B. über einem Link auf die digitale Ressource oder den Eintrag der Lehrveranstaltung im jeweiligen Vorlesungsverzeichnis).

26 http://www.edulap.ch

Das Orientierungssystem von edulap ist kein Repository, sondern eine Metadatenplattform. edulap vereint die derzeit nur dezentral verfügbaren (virtuellen) Lehrangebote unter einem Dach und schafft erstmalig einen zentralisierten Überblick über das Lehrangebot und die damit verbundenen, zahlreich vorhandenen Unterrichtsmaterialien, die für das Selbststudium oder aber auch als Ergänzung des Präsenzunterrichts genutzt werden können. Das am Beispiel des Massenfaches Psychologie entwickelte Konzept wird zu einem fortgeschrittenen Projektzeitpunkt unmittelbar auf andere Fächer übertragbar sein.

In diesem Kapitel soll untersucht werden, ob verschiedene Experten die Lehritems in einer konsistenten Weise anordnen und falls ja, ob die HM vergleichbare Konfigurationen produziert. Dazu erstellen wir eine Expertenkarte und prokrusten sie mit dem Output der HM.

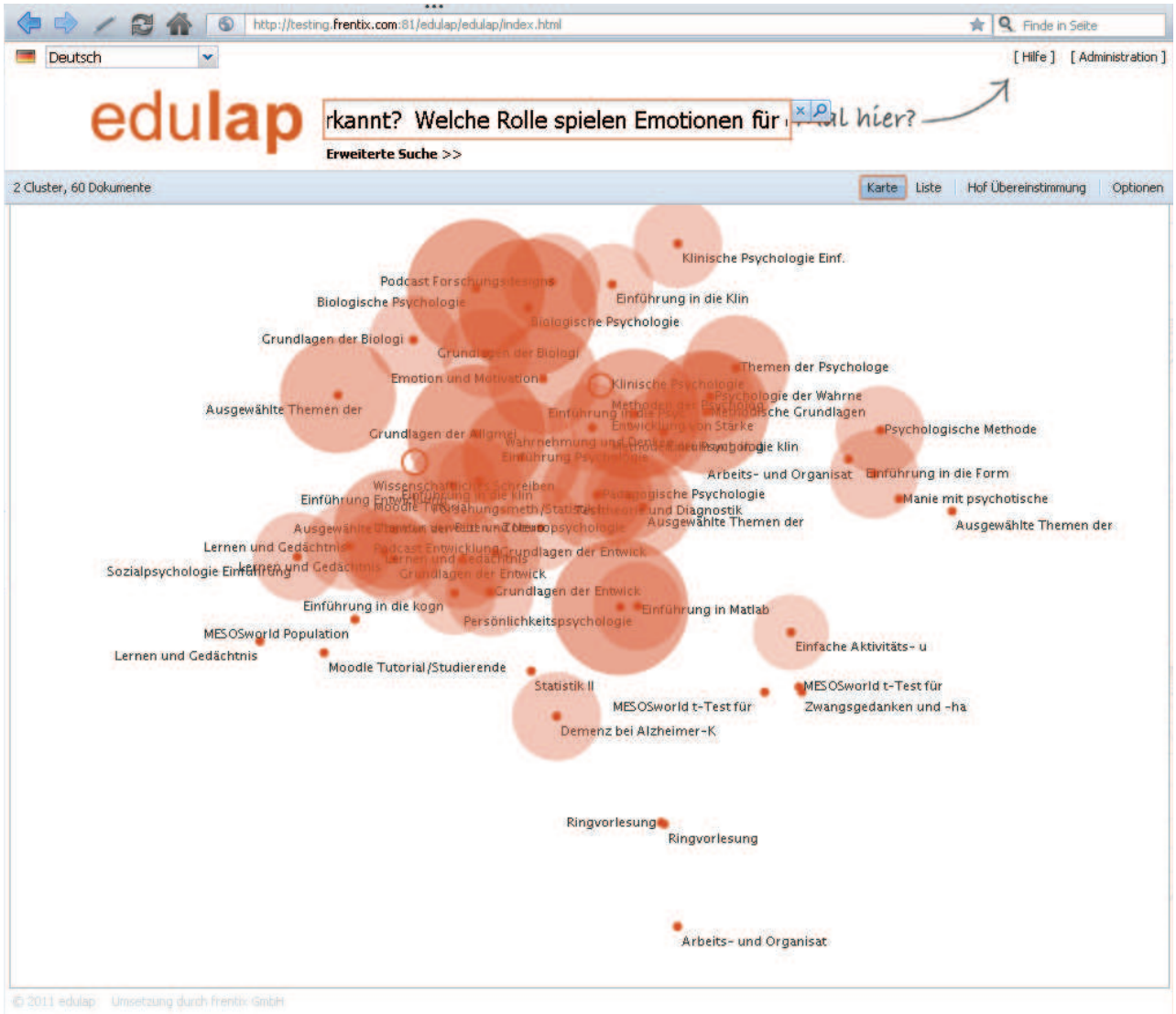


Abbildung 91: Screenshot einer Karte aus der laufenden Testumgebung (Stand Mai 2012). Die Resultate einer Suche sind in einer semantischen Karte dargestellt. Zusätzlich sind in diesem Beispiel die semantisch ähnlichsten Bereiche eingefärbt.

23.3 Vorgehen

Aus dem Datenbestand des edulap-Projekts wird eine zufällige Auswahl von 70 deutschsprachigen Items getroffen. Sechs Versuchspersonen (alles Mitarbeitende von edulap) durchlaufen damit zwei Sessions. In der ersten müssen sie jeweils ein paralleles und ein hierarchisches Sortieren durchführen; in der zweiten Session – ein paar Tage später – nochmals ein paralleles Sortieren (zwei VPn nahmen an dieser Wiederholung nicht teil). Beide Sortierarten fanden am Computer statt. Da dort nur die Titel der Items ersichtlich waren, wurden den VPn die Titel und Beschreibungen (Abstracts) der 70 Items zusätzlich ausgedruckt und beigelegt.

Diese Stichprobe von sechs VPn genügt knapp den statistischen Kriterien, um stabile NMDS-Karten zu erzeugen, sie ist aber nicht repräsentativ, weder was das Zielpublikum anbelangt (was in diesem Fall auch nicht beabsichtigt ist), noch was das Expertentum betrifft. Es soll hier keine globale Expertenkarte geeicht werden, sondern eine Vorgabe gebaut werden, die als Wegweiser für die HM dient.

Dieselben 70 Items werden anschliessend behoft, wobei die Keywords aus der Regensburger Verbundklassifikation[27] stammen. Die Normierung wird gemäss TotalTargetWords erstellt, das Hofgewicht auf 10 festgelegt, mit ÜK (Gewicht 1) und ohne Kategorienkonstante. Ein Item (Visuelle Wahrnehmung) wurde in den HM-Karten nicht berücksichtigt, da das Abstract sowohl in der Länge als auch in der Qualität ungenügend war.

Schliesslich wird die Expertenkarte mit der Hofkarte prokrustet, um die Verschiedenheiten herauszuarbeiten.

23.3.1 Das parallele Sortieren (PS)

Beim PS verteilt die VP sämtliche Items nach eigenen Kriterien in eine selbst zu bestimmende Anzahl Kategorien. Bei der verwendeten Variante müssen diese Kategorien nicht benannt werden. Anschliessend wird eine Dreiecksmatrix erstellt, in der das gemeinsame Vorkommen in einer Kategorie zwischen zwei Items mit einer 1 eingetragen wird, beziehungsweise im negativen Fall mit einer 0 (s. Abb. 92).

27 http://www.uni-regensburg.de/bibliothek/

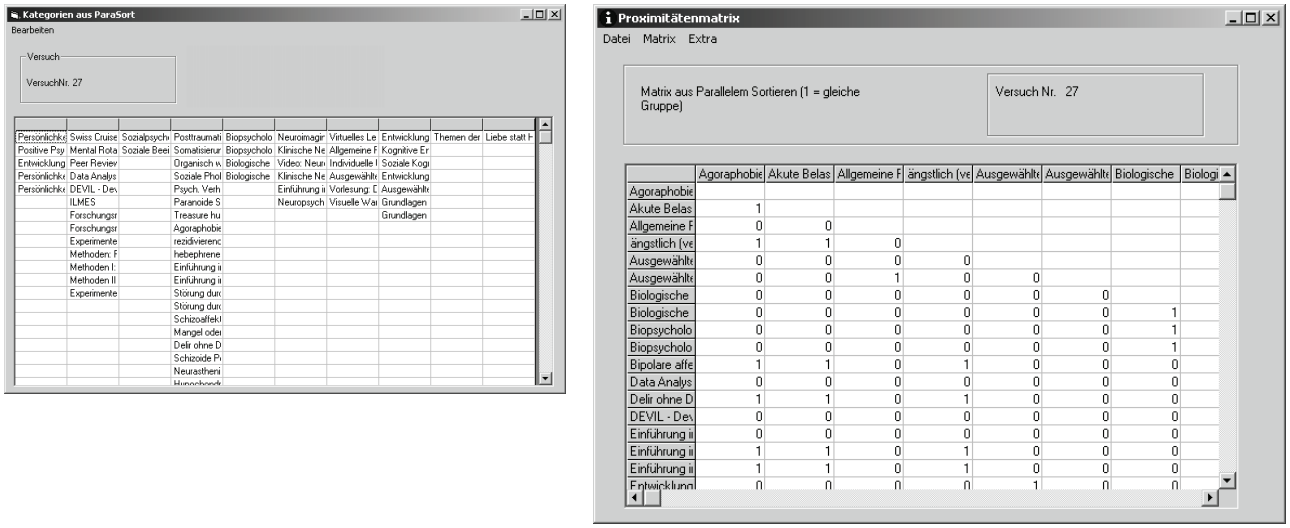


Abbildung 92: Links ist die erfolgte Kategorieneinteilung nach dem PS einer VP zu sehen. Rechts ist die dazugehörige Dreiecksmatrix. Eine «1» zeigt das gemeinsame Auftreten dieser Items innerhalb einer Kategorie an, eine «0» dessen Fehlen.

Das PS dauert relativ kurze Zeit (ca. 15 Minuten). Es hat den Vorteil, dass es zeigt, was für eine grobe Strukturierung eine VP macht. Durch die Wiederholung einige Tage später zeigt sich eine gewisse Durchlässigkeit der Kategorien. Wenn eine VP ein Item das erste Mal in Kategorie A einordnet, beim zweiten Mal aber in Kategorie B, ist das ein Indiz dafür, dass die Grenzen zwischen diesen Kategorien nicht in allen Bereichen scharf sind.

23.3.2 Das hierarchische Sortieren (HS)

Beim HS werden die Items von der VP zuerst in zwei Gruppen geteilt. Wiederum ist das Sortierkriterium völlig frei. Danach müssen die Items der ersten Gruppe in zwei Untergruppen aufgeteilt werden, anschliessend die Items der zweiten Gruppe, usw. Das wird so lange gemacht, bis nur noch ein, beziehungsweise zwei Items pro Gruppe übrig bleiben. Das HS dauert ca. 45 Minuten und wird von den VPn als eher mühsam empfunden, da das Aufteilungskriterium gegen den Schluss hin praktisch arbiträr wird.

Der Vorteil dieses Verfahrens liegt darin, dass eine Intraclusterstruktur eruiert werden kann, die Aufteilung ist gewissermassen feinmaschiger. Dementsprechend werden NMDS-Karte weniger deutlich geclustert.

Auch beim HS wird eine Dreiecksmatrix erstellt: Es wird gezählt, über wie viele Hierarchiestufen man von Item A zu Item B kommt. Diese absoluten Zahlen werden zwischen 0 und 1 normiert und der Wert von 1 subtrahiert, damit hohe Zahlen hohe Ähnlichkeiten ausdrücken (s. Abb. 93) und die Matrix somit direkt vergleichbar mit derjenigen aus dem PS ist.

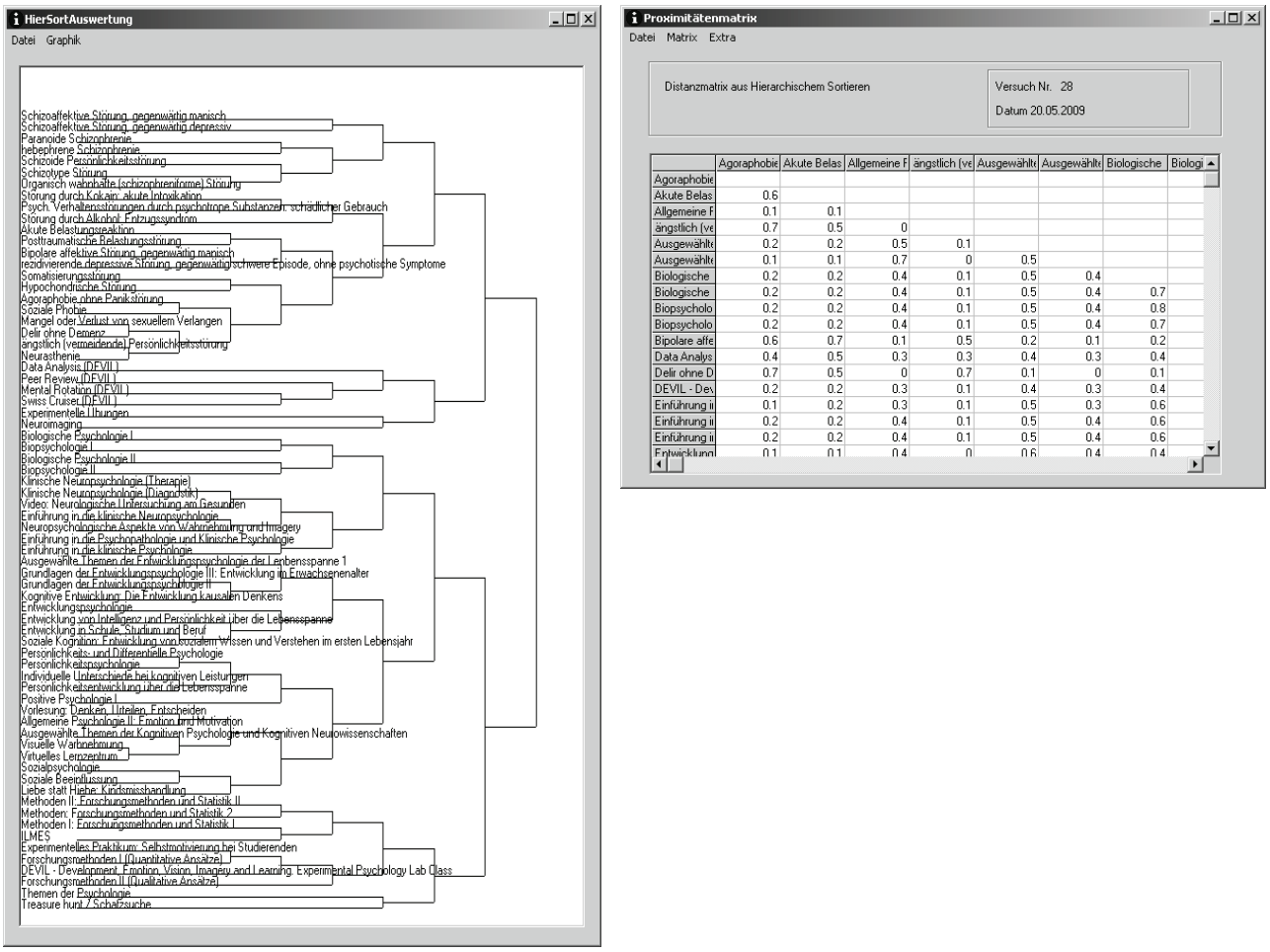


Abbildung 93: Links ist die erfolgte Strukturierung aller Items nach dem HS einer einzigen VP sichtbar. Rechts ist ein Ausschnitt der dazugehörigen Dreiecksmatrix.

23.4 Resultate

23.4.1 Die Expertenkarte

Die 70 Items setzten sich vor allem aus Beschreibungen von psychischen Störungen (aus der Online-Lernumgebung PTO[28]) und verschiedenen, psychologischen Vorlesungen zusammen, sowie einzelnen

28 http://www.psychopathology.ch/

Abweichungen (ein Video, eine Online-Plattform, ein Online-Spiel). Die Vorlesungen waren quer über die Themenbereiche der Psychologie verstreut, jedoch mit einem hohen Anteil methoden- und statistik-bezogener Texte.

Diese Grundstruktur findet sich nun auch bei allen Variationen der errechneten Karten: Bei den einzelnen VPn beim PS, wie auch beim HS, sowie – demzufolge – bei den aggregierten Daten. Ein prototypisches Beispiel aus dem PS zeigt Abbildung 94. Die VP unterteilte die Texte scharf in psychische Störungen, Statistik-/Methodentexte und restliche Texte. Im (schwarzen) Cluster der psychischen Störungen befinden sich auch einige (blauen) Texte aus der klinischen Psychologie, z. B. «Einführung in die Psychopathologie und Klinische Psychologie», «Einführung in die klinische Psychologie» und «Treasure hunt / Schatzsuche». Die drei blauen Texte (einer ist verdeckt) im Cluster der restlichen Texte betreffen alle die Neuropsychologie, welche die VP offenbar nicht im klinischen Bereich einordnen würde, wie es unsere Einfärbung suggeriert.

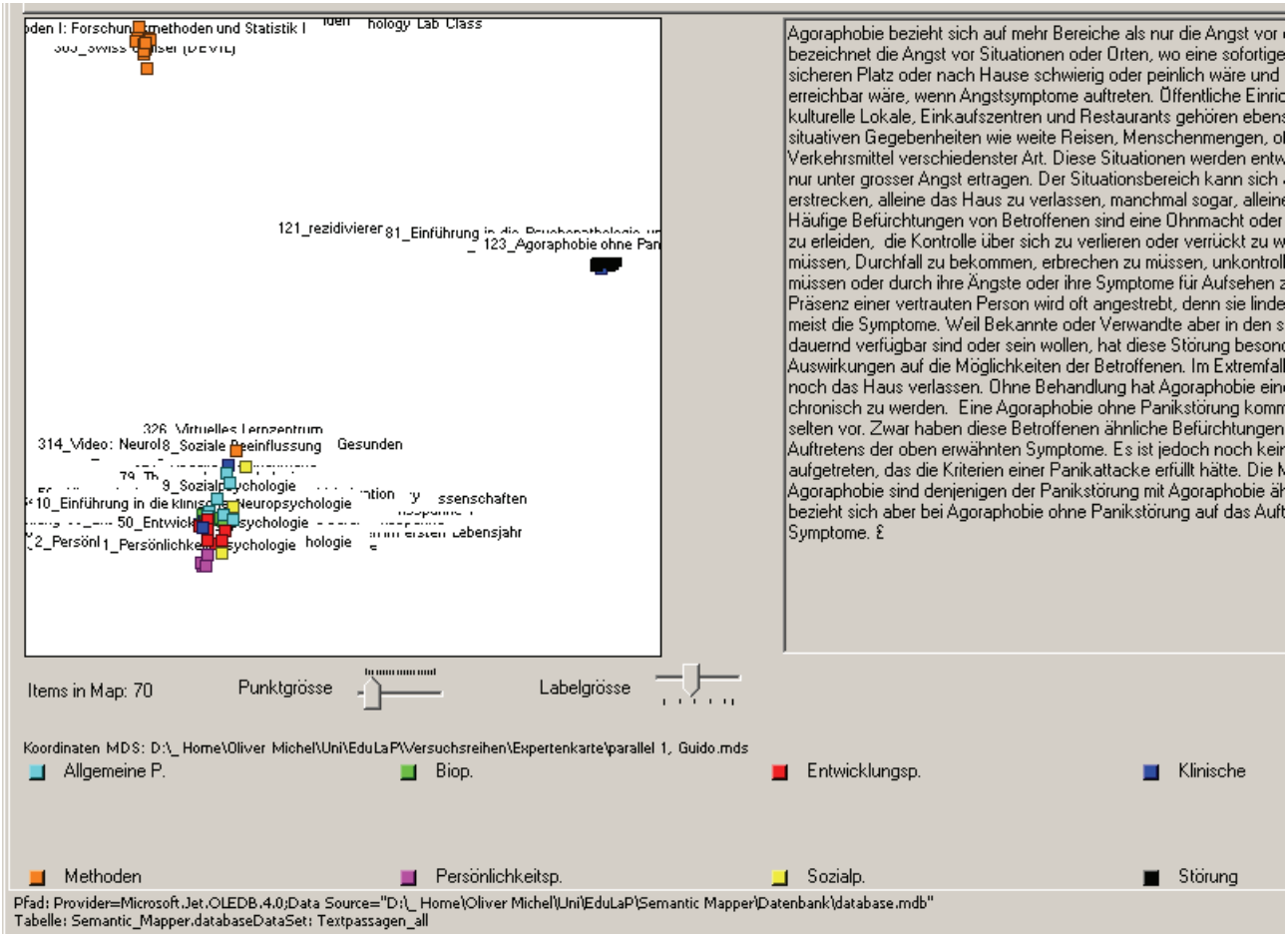


Abbildung 94: Typische 3er-Clusterung aus dem parallelen Sortieren: Die Störungsbilder, die methodischen Vorlesungen und die übrigen Vorlesungen bilden jeweils eine Kategorie.

Die Karten aus dem ersten und zweiten PS zeigen keine grossen Unterschiede: Zuerst wurden die Matrizen aus dem ersten PS gemittelt, ebenso diejenigen aus dem zweiten PS, anschliessend wurden die beiden gemittelten Karten prokrustet. Die Prokrustes-Transformation macht deutlich, dass sich die Grundstruktur nicht ändert, aber die «Exoten» unter den Items springen (s. Abb. 95).

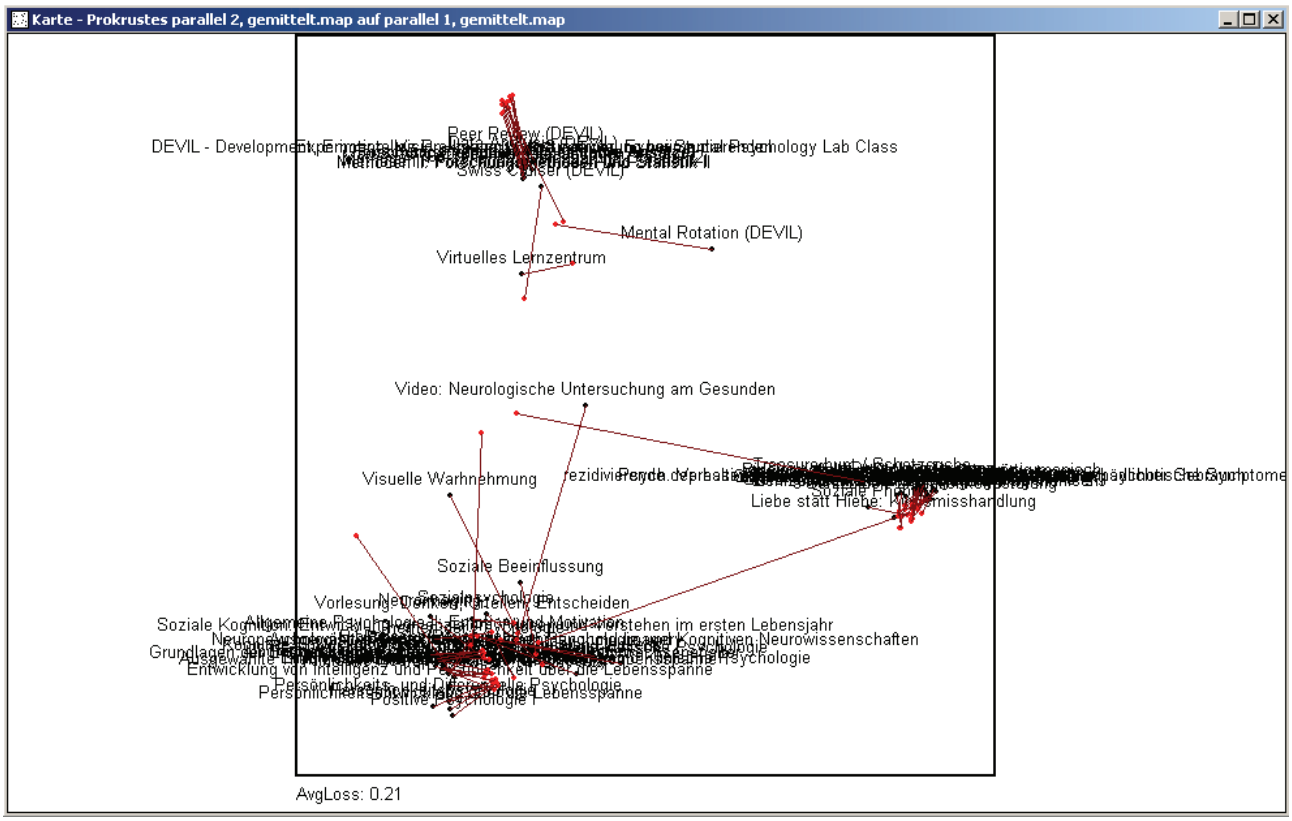


Abbildung 95: Prokrustes-Transformation der gemittelten Daten aus dem ersten und zweiten PS: Die Grundstruktur bleibt sich gleich, aber die exotischen Items springen.

Die unschärfere Clusterung des HS zeigt sich in Abbildung 96. Die Grundstruktur der drei Kategorien ist nach wie vor deutlich, jedoch gibt es fliessende Übergänge. Im Zentrum der exemplarischen Karte sind die Kategoriezuordnungen schliesslich nicht mehr eindeutig. So sind beispielsweise die klinischen Vorlesungen (blau) zwischen den übrigen Vorlesungen und den psychischen Störungsbildern verortet.

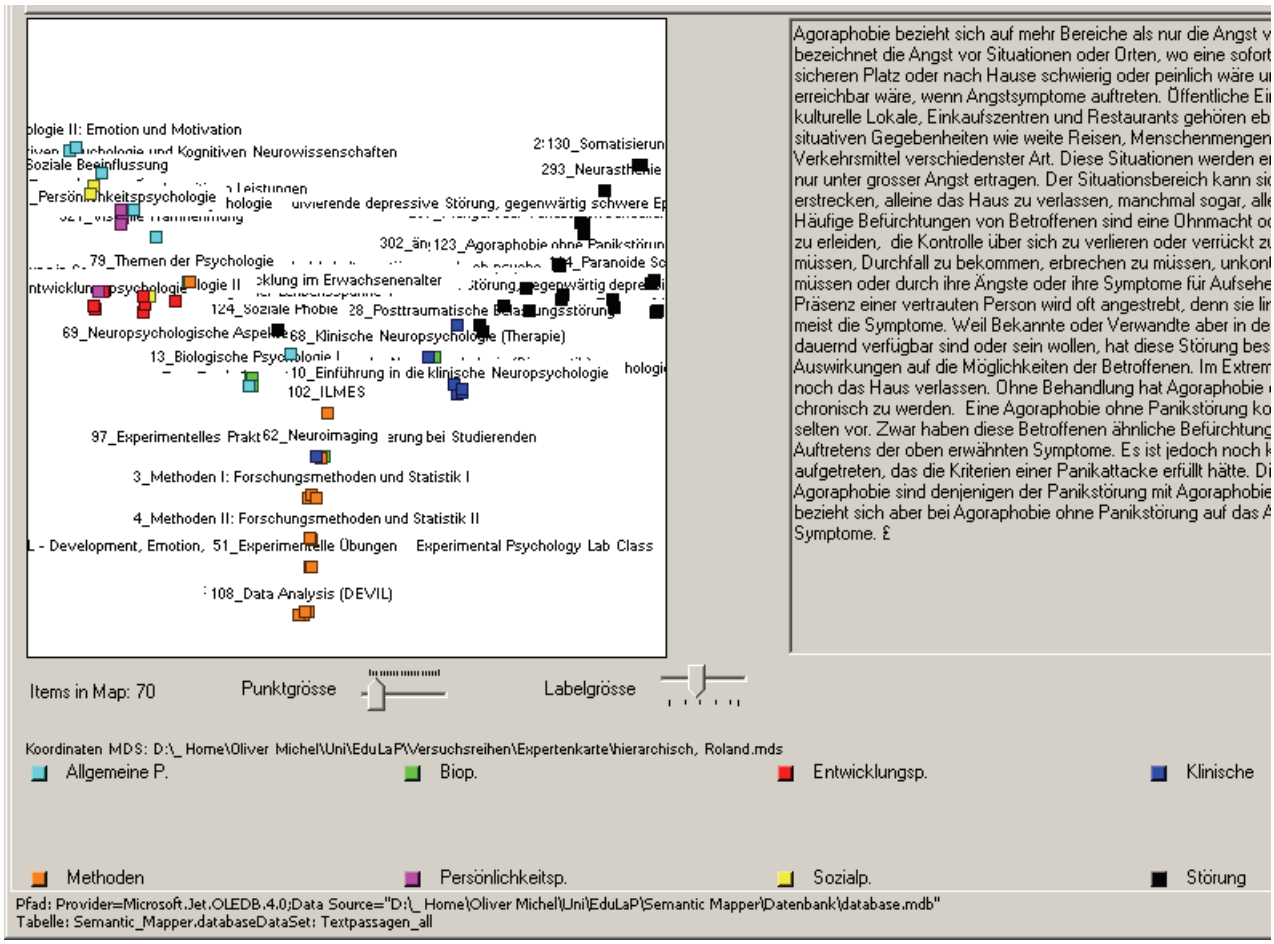


Abbildung 96: Ein Einzelbeispiel aus dem HS: Die Clusterung ist weniger deutlich, die grobe Struktur bleibt aber derjenigen aus dem PS ähnlich.

Legt man die gemittelten Daten aus dem ersten und zweiten PS auf die gemittelten Daten aus dem HS, zeigt sich als Bild, was in Worten schon gesagt wurde: Beim HS wird die harte Clusterung etwas aufgeweicht, die Grundstruktur bleibt aber erhalten (s. Abb. 97).

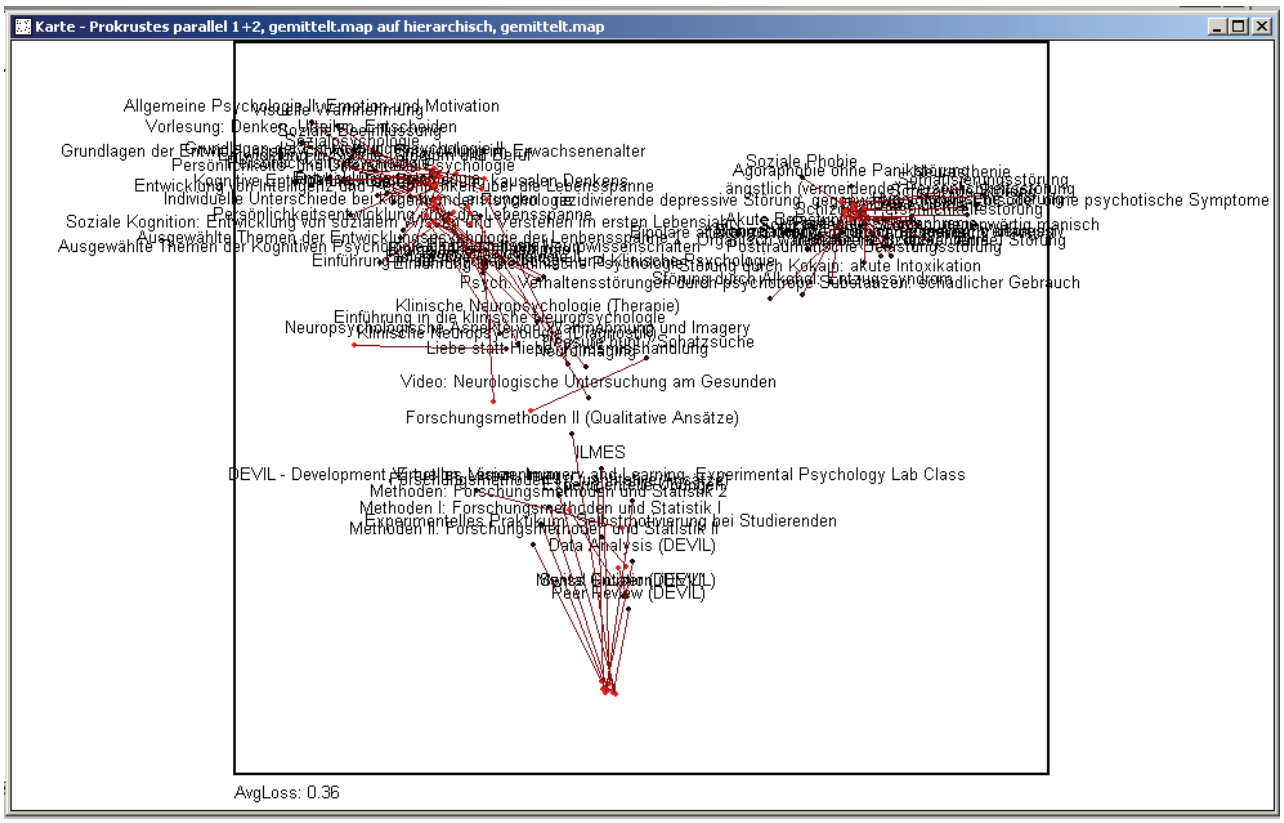


Abbildung 97: Prokrustes-Transformation der gemittelten Daten aus dem ersten und zweiten PS, sowie dem HS: Die Grundstruktur bleibt erhalten. Die enge Clusterung aus dem PS (rote Punkte) wird etwas aufgeweicht.

Die Expertenkarte ergibt sich nun aus der Mittelung der Daten aus dem ersten und zweiten PS einerseits und dem HS andererseits (s. Abb. 98).

Die Expertenkarte lässt sich folgendermassen charakterisieren: Drei deutliche Cluster spannen das Feld auf. Diese Cluster bestehen aus «Störungsbildern», «statistikbezogener Vorlesungen» und «restliche Vorlesungen». Dazwischen liegen exotische Lernitems (eine DVD über Kindsmisshandlung, ein Video über neurologische Untersuchungen, ein verhaltenstherapeutisches Computerspiel), aber auch inhaltliche Mischformen. So liegt «Klinische Neuropsychologie (Diagnostik)» zwischen dem Cluster «übrige Vorlesungen» und «Statistik», ebenso «Themen der Psychologie» und «Neuroimaging», «Einführung in die klinische Psychologie» und «Einführung in die Psychopathologie und Klinische Psychologie» liegen fast aufeinander und zwischen «übrige Vorlesungen» und «Störungen».

Unpassende Einordnungen gibt es keine. Höchstens das «Virtuelle Lernzentrum», welches im Statistik-Cluster liegt, aber genauso gut zwischen diesem und den übrigen Vorlesungen sein könnte.

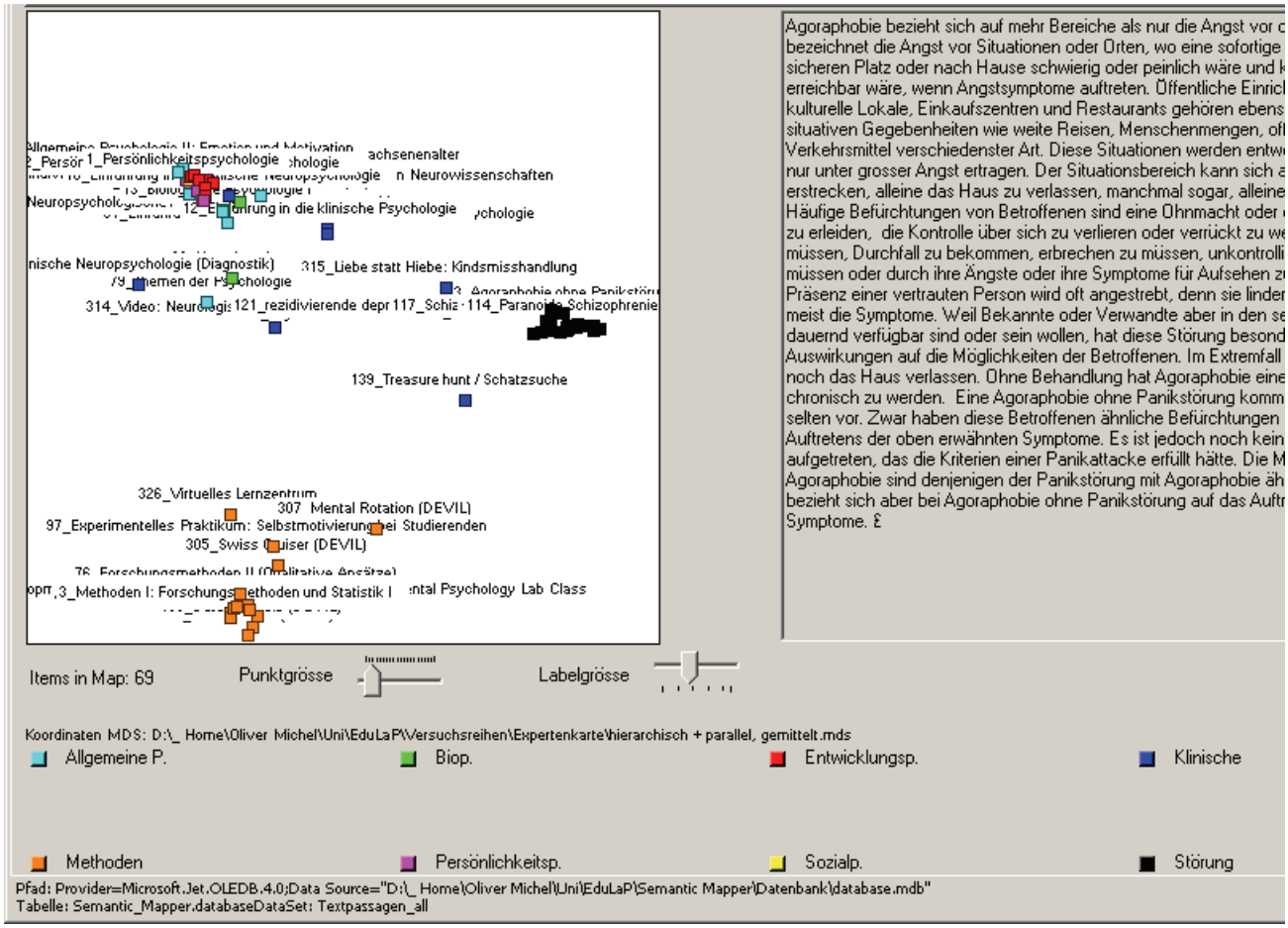


Abbildung 98: Die Expertenkarte: Mittelung aus den Daten aus dem ersten und zweiten PS, sowie dem HS.

Abbildung 99 zeigt wieder die Expertenkarte, diesmal wurden die einzelnen Texte aber nach den drei Grundkategorien eingefärbt, wobei die Kategoriezuweisung nicht aufgrund der Karte erfolgte, sondern nach Inhalt. Dass diese Kategorisierung so überaus deutlich in der Karte widerspiegelt wird, ist nicht zwingend.

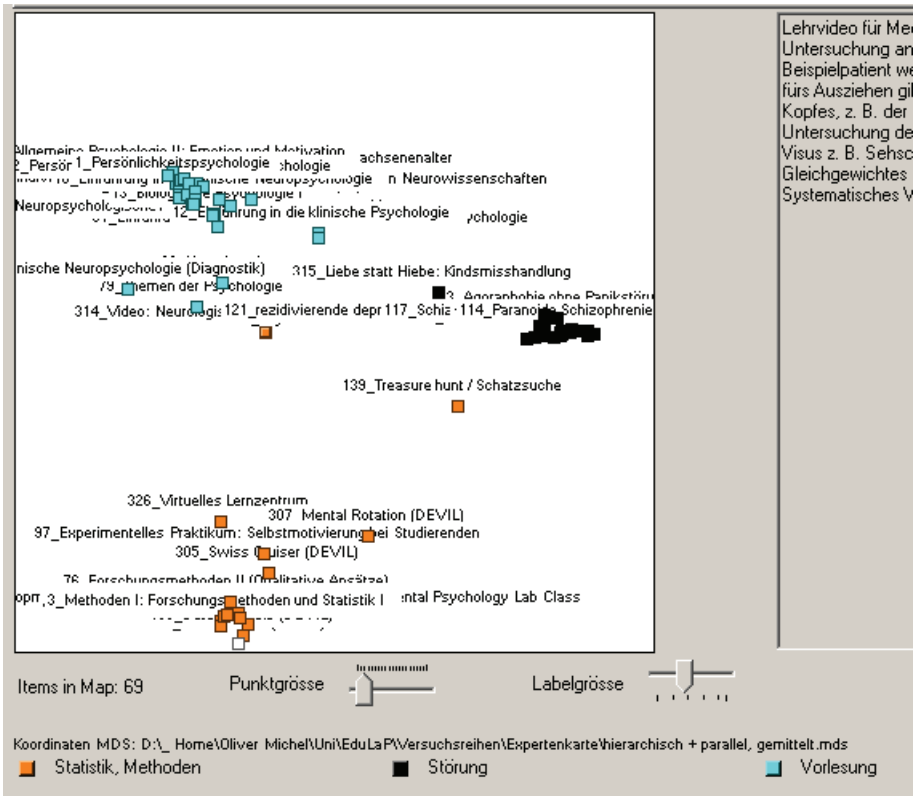


Abbildung 99: Die Expertenkarte, eingefärbt nach den drei Grundkategorien

23.4.2 Vergleich mit der Hofmethode

Obwohl das hier verwendete Datenmaterial recht heterogen ist, produziert die Hofmethode eine Karte, die in der Gesamtstruktur mit der Expertenkarte übereinstimmt. In Abb. 100 sind wieder die drei Bereiche Statistik, Störungsbilder und übrige Vorlesungen erkennbar. Überaus verblüffend ist die Tatsache, dass die drei Exoten «314, Video: Neurologische Untersuchung am Gesunden», «139, Treasure hunt / Schatzsuche» und «315, Liebe statt Hiebe: Kindsmisshandlung» - die in der nach den Grobkategorien eingefärbten Expertenkarte (Abb. 99) alle ausserhalb ihrer Cluster stehen – auch in der HM-Karte zwischen die Bereiche zu liegen kommen. In Abbildung 101 ist die HM-Karte nochmals dargestellt, dieses Mal jedoch nach den Grobkategorien eingefärbt und mit Markierung der drei Exoten.

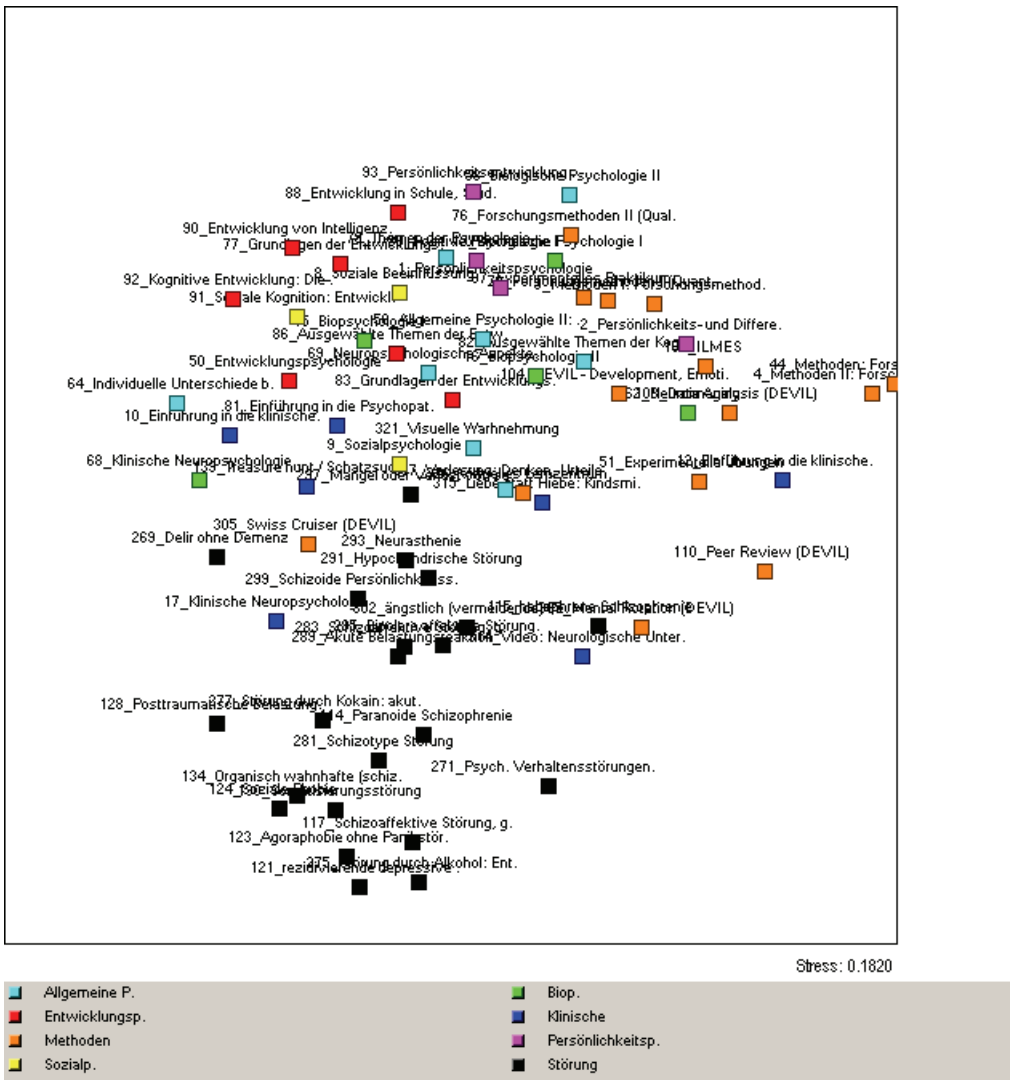


Abbildung 100: Semantische Karte basierend auf der Hofmethode; wieder sind die Grundstrukturen erkennbar, wenn auch ohne Clusterung.

Offensichtlich falsche Platzierungen gibt es sehr wenige. Der unscharfe Übergang zwischen den Störungsbildern und den übrigen Vorlesungen besteht mehrheitlich aus klinischen Texten, was inhaltlich gewünscht ist. Einzig das Item «305, Swiss Cruiser (DEVIL)» wäre eher im kognitiven Bereich der Vorlesungen zu vermuten, nicht im klinischen.

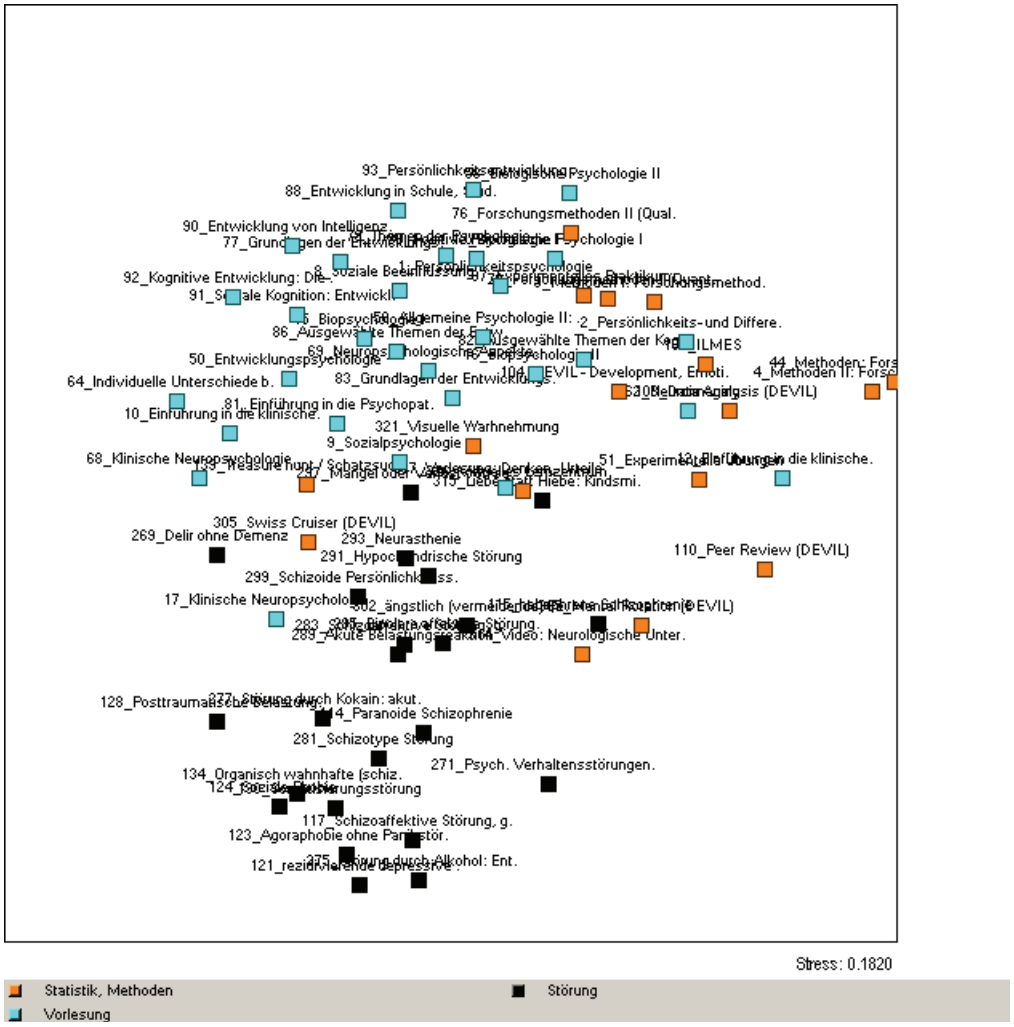


Abbildung 101: Die HM-Karte eingefärbt nach den Grobkategorien. Hervorgehoben sind die drei Exoten, die auch in der Expertenkarte zwischen den Kategorien platziert sind.

Auch die Prokrustes-Transformation (Abb. 102) macht deutlich, wie wenig sich die drei Grobkategorien unterscheiden: Bei der HM-Karte sind allerdings die Grenzen fliessend, während die Expertenkarte deutlich clustert.

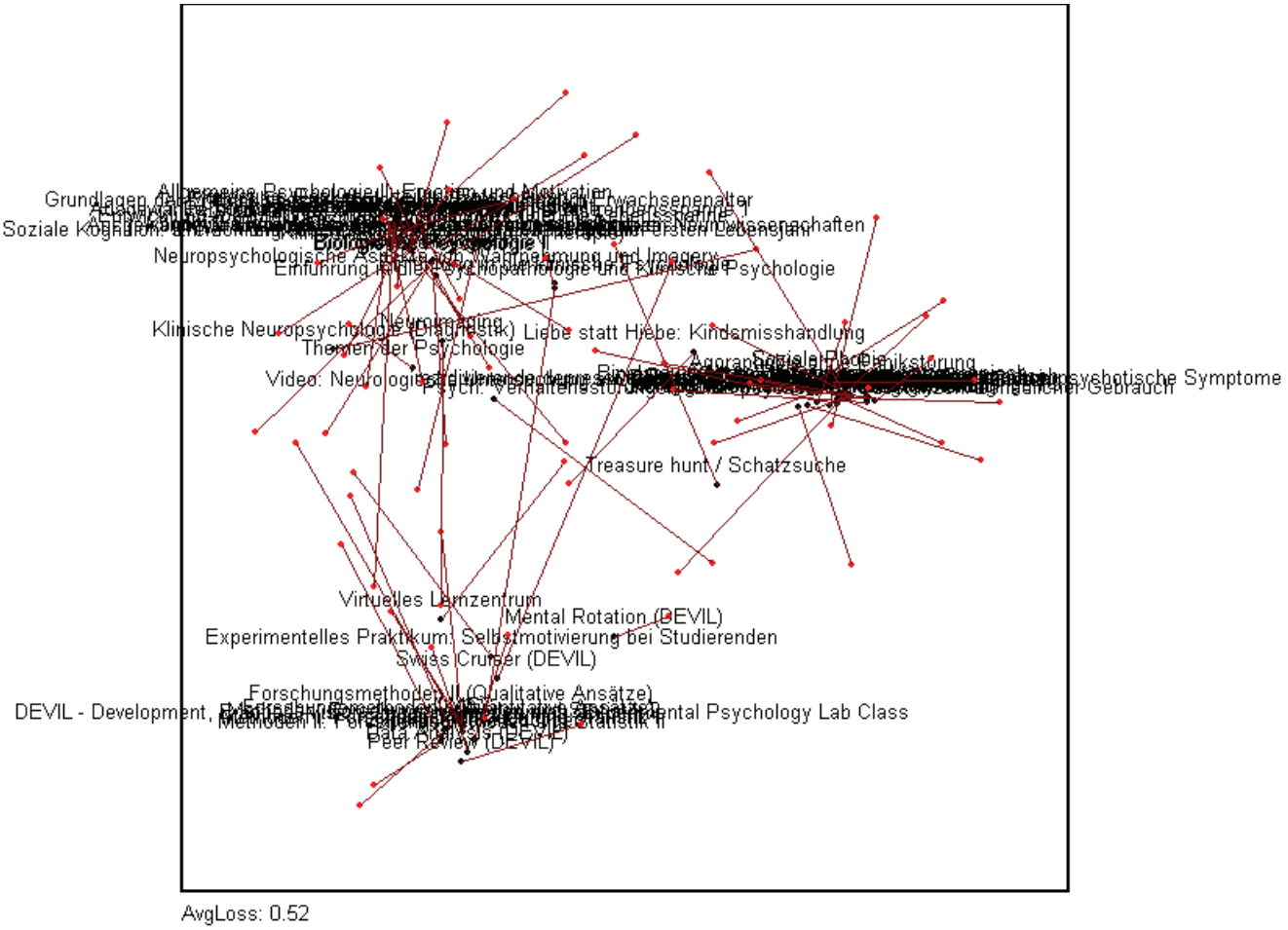


Abbildung 102: Prokrustes-Transformation der Expertenkarte (schwarz) mit der HM-Karte (rot). Die drei Grobkategorien sind bei beiden Karten vorhanden und durchmischen sich nur wenig.

23.5 Diskussion

Die Expertenkarte scheint eine inhaltlich vernünftige Vorgabe für die Hofmethode zu liefern. Die drei resultierenden Hauptstrukturen sind inhaltlich nachvollziehbar und konsistent.

Die Karte der Hofmethode wiederholt nicht nur diese Grundstrukturen, sondern repliziert ebenfalls diejenigen Platzierungen von Texten, welche schon bei der Expertenkarte uneindeutig waren.

Somit kann anhand dieses explorativen Experimentes festgehalten werden, dass die Hofmethode mit den Texten des edulap-Projekts grundsätzlich gut funktioniert. Wie sich das auf die Usability der Karten auswirkt, muss noch untersucht werden.

24 Wikipedia-Experiment

24.1 Überblick

Im Wikipedia-Experiment werden Texte aus dem Online-Lexikon Wikipedia einerseits mit der Hofmethode, andererseits mit dem Überlappungskoeffizienten strukturiert und Versuchspersonen vorgelegt. Es zeigt sich, dass die Berechnungsgrundlage zweitrangig ist; wichtiger ist die Lage eines zu findenden Zielitems. Versuchspersonen schätzen Strukturierungen mit Clustern. Es zeigt sich ebenfalls ein technisches Erfordernis: Es muss sichergestellt werden, dass Zielitems nicht durch Nachbarn überdeckt werden.

24.2 Einleitung

Im Wikipedia-Experiment werden drei Fragen untersucht. Erstens interessiert, wie sich die Hofmethode in Bezug auf die Strukturierung von Texten im Vergleich zum Überlappungskoeffizienten verhält. Als Datengrundlage dient die Wikipedia[29]. Diese Datenbank stellt einen gewaltigen Fundus an breit gefächerten, aus der Praxis generierten Thementexten bereit.

Zweitens wird ein HM-spezifischer Aspekt untersucht, nämlich die Stichwortauswahl. So stellt sich die Frage: Wenn man eine Textauswahl aufgrund einer konventionellen Suche und mit verschiedenen Stichworten gefunden hat, nach welchem/welchen Stichwort/-wörtern sollen diese Texte anschliessend strukturiert (behoft) werden, damit der gesuchte Lösungstext möglichst rasch gefunden wird?

Drittens werden zwei verschiedene Aufgabenarten untersucht: Zielgerichtete und offene Fragen. Bei den zielgerichteten Fragen kommt nur ein ganz bestimmter Text als Lösung in Frage; bei den offenen Frage gibt es viele Möglichkeiten, von denen ein Teil gefunden werden muss. Bei welchen Fragen eignet sich der Einsatz der HM?

29 http://www.wikipedia.org

24.3 Vorgehen

Um die semantische Strukturierung der Karten zu beurteilen, erstellten wir verschiedene Karten (einerseits basierend auf der HM – dabei wurden unterschiedliche Stichworte behoft – andererseits auf dem ÜK) und liessen VPn Fragen beantworten, die allgemeiner Natur oder spezifisch auf einzelne Stichworte gerichtet waren.

Hypothesen

Was ist die qualitative Leistung der HM im Vergleich zum ÜK in der Anwendung auf reale Texte?

1. Hypothese: Die Strukturierung von Texten in einer Karte hilft bei der Textsuche (verglichen mit einer Zufallssuche in der Karte). Das betrifft die HM und den ÜK.
2. Hypothese: Die aufgabenspezifische Strukturierung (HM) ordnet semantisch besser als die allgemeine Strukturierung (ÜK) und hilft demzufolge besser bei der Aufgabenlösung.

Als Textkorpus diente die deutschsprachige Ausgabe der Wikipedia. Die gesamte Datenbank wurde auf den lokalen Computer gespiegelt und mit der Open-Source-Suchmaschine Lucene[30] indiziert. Wir durchsuchten dann die Wikipedia nach Texten, in denen die Wörter «Luft», «Wasser» und mindestens eines der drei Wörter «Erde, Boden, Grund» vorkamen (Boolescher Suchausdruck: Luft AND Wasser AND [Erde OR Boden OR Grund]). Die drei Wörter Erde/Boden/Grund können synonym gebraucht werden. Im Gegensatz zur klassischen Suche sollen sie die Resultatmenge aber nicht verwässern, sondern die HM soll den synonymen Gebrauch – die semantische Schnittmenge – darstellen, quasi zeigen «was der User eigentlich meint». Insgesamt 76 Texte entsprachen diesen Kriterien.

Wir erstellten fünf verschiedene Kartentypen:

- drei Karten basierend auf den Hofähnlichkeiten eines der Stichwörter «Luft», «Wasser» und «Erde, Boden, Grund»
- eine Karte mit allen drei Stichwortgruppen
- eine Karte basierend auf dem ÜK

Diese Karten punktspiegelten wir zusätzlich, damit wir den VPn eine grössere Auswahl von Karten vorlegen konnten, ohne einen Wiedererkennungseffekt zu riskieren. Insgesamt hatten wir demnach 10 verschiedene Karten.

30 http://lucene.apache.org/

Die VPn mussten 10 Suchaufgaben lösen. Die Rahmengeschichte handelte davon, dass die VP eine Lehrperson sei, die mit ihrer Klasse eine Projektwoche zu den Elementen Luft, Wasser und Erde durchführen wolle und eine Internetrecherche durchgeführt habe; nun habe sie sich einige Fragen aufgeschrieben und wolle diese mit den gefundenen Resultaten beantworten.

Die Suchfragen (s. Anhang 5) bestanden aus verschiedenen Kategorien (in Klammern ist die interne Nummerierung angegeben):

- vier allgemeine Fragen (0–3)
- zwei Wasser-spezifische Fragen (4, 5)
- zwei Luft-spezifische Fragen (6, 7)
- zwei Fragen des Typs «3 aus vielen» (8, 9)

Die allgemeinen Fragen liessen sich nicht einer bestimmten Elementkategorie zuordnen, z.B. «Welcher Beruf beschäftigt sich mit dem Thema?».

Die spezifischen Fragen nahmen explizit Bezug auf ein Element, z. B. «Wie kann man angeblich die Raumluft reinigen?». Der Typ «3 aus vielen» bedeutet, dass zur Frage mehrere (ca. 10) Lösungstexte passten, jedoch nur 3 davon gefunden werden mussten. Dieser Fragetyp hatte den willkommenen Nebeneffekt die VPn zu motivieren: Bei früheren Arbeiten hatte sich gezeigt, dass diese Texte, solange relativ eng geclustert, gut gefunden und abgeklickt werden konnten, was den VPn das angenehme Gefühl gab, die Karte begriffen zu haben.

Die Frage 0 besitzt zwei Lösungsitems, die Fragen 1–7 je eine Lösung, die Fragen 8 und 9 zehn, beziehungsweise elf mögliche Lösungsitems.

Die Fragen-Karten-Kombinationen wurden ausgewogen verteilt und die Reihenfolge randomisiert. 30 VPn wurden getestet, was 6 Werte pro Frage-Karte-Kombination ergibt. Den VPn wurde jeweils eine Frage und eine Karte präsentiert, in der die Texte durch kleine Punkte repräsentiert wurden. Durch Anklicken dieser Punkte erschien in einem anderen Bereich des Bildschirms der zu Grunde liegende Text. Es wurde sofort erkenntlich gemacht, ob es sich dabei um einen gesuchten Lösungstext handelte. Erhoben wurde die Anzahl Klicks, die die VPn brauchten, um den/die Lösungstext/-e zu finden, wobei sie sich bereits angeklickte Items beliebig oft nochmals anschauen konnten, ohne dass diese Klicks gezählt wurden.

24.4 Resultate: Kartenfaktoren

Im Mittel wurde jede Aufgabe mit 28.4 Klicks gelöst. Die Mittelwert-Unterschiede zwischen den Karten sind – über alle Fragen hinweg zusammengefasst – nicht signifikant (s. Abb. 103). Das heisst, dass die zweite Hypothese (HM-Karten strukturieren besser als ÜK-Karten) verworfen werden muss.

Wir teilten die Fragen-Karten-Kombinationen in Kategorien ein, die nach der Lage der Lösungsitems unterschieden, rechneten die MW aus und machen einen T-Test. Es ist in Abbildung 104 ersichtlich, dass es ein leichter Vorteil ist, wenn sich das Lösungsitem innerhalb eines Clusters befindet, sogar dann, wenn es verdeckt wird.

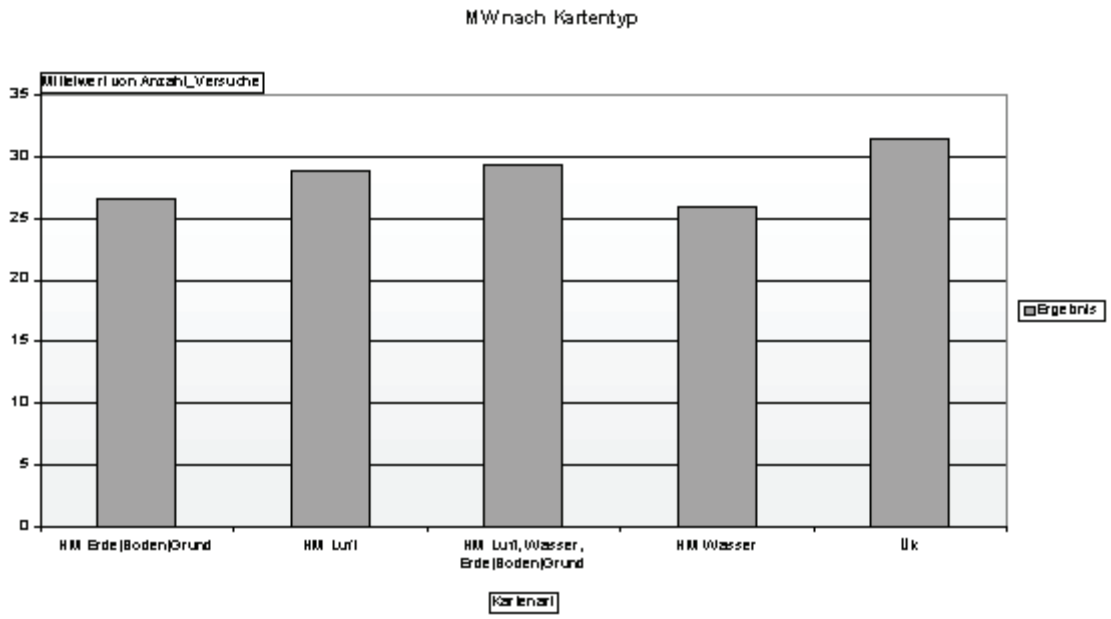


Abbildung 103: Die MW-Unterschiede zwischen den Karten werden über alle Fragen hinweg zusammengefasst nicht signifikant.

Interessanter sind aber die einzelnen Fragen-Karten-Kombinationen. Gibt es bestimmte Fragen, die sich mit bestimmten Karten besser oder schlechter lösen lassen als mit anderen und warum?

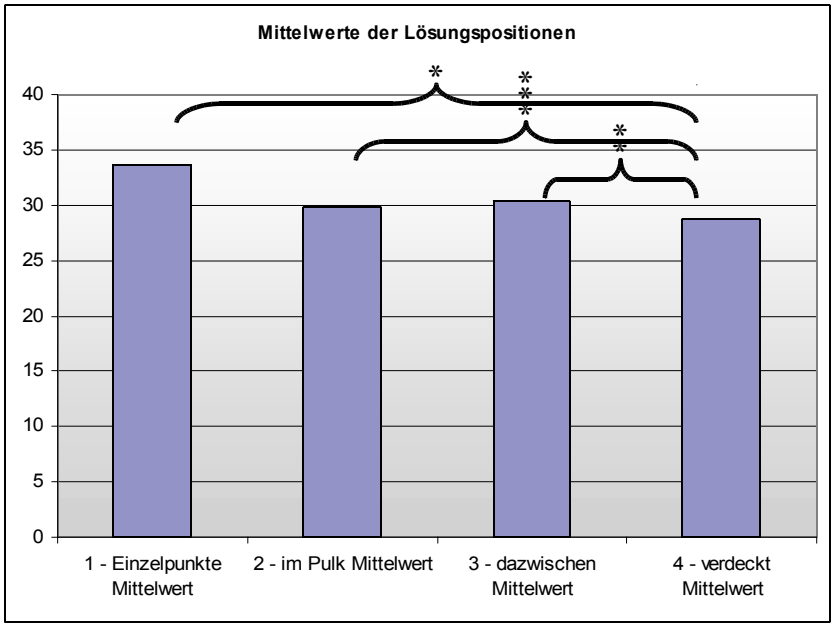


Abbildung 104: Die Position der Lösungsitems spielt über das Ganze eine nicht zu unterschätzende Rolle; tendenziell ist es besser, wenn sich das Lösungsitem in einem Cluster befindet, als wenn es separat platziert ist.

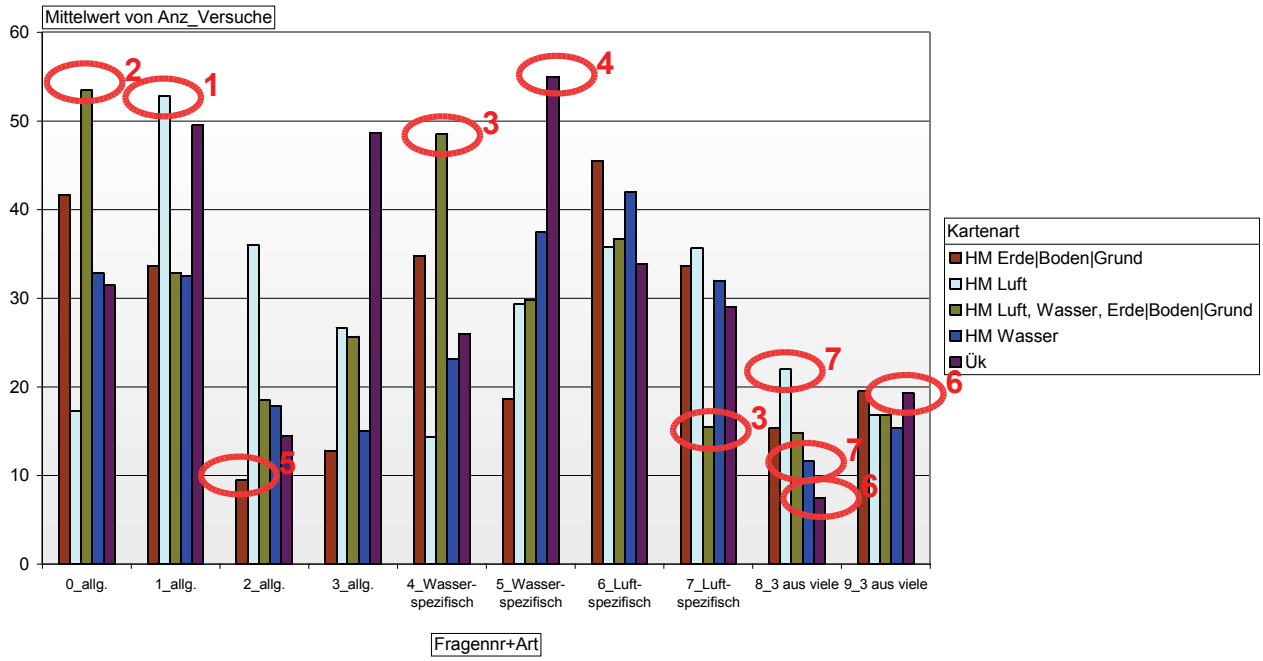


Abbildung 105: Auf der X-Achse sind die 10 Fragen aufgereiht, auf der Y-Achse die Mittelwerte (basierend auf jeweils sechs Zahlen). Besonders gute oder schlechte Fragen-Kartenkonstellationen sind eingezeichnet.

24.4.1 Prägnante Konstellationen

Abbildung 105 zeigt die gemittelten Klickraten aller Fragen, aufgeteilt auf die einzelnen Karten (die gespiegelten Karten werden nicht separat ausgewiesen). Aussergewöhnliche Konstellationen sind eingezeichnet und werden besprochen.

Besonders schlechte Konstellationen:

1. **Frage 1 mit Karte 0/5 (Luft), MW 53, Abb. 106**
   Durchschnittlich 53 Versuche brauchten die drei VPn (bei total 76 Items!). Das Zielitem ist in der Abbildung 106 grün eingefärbt, jedoch nicht sichtbar, da es durch andere Items vollständig überdeckt wird. Dieser triviale Umstand darf für konkrete Anwendungen der semantischen Karten nicht vergessen gehen: Es muss verhindert werden, dass sich zu viele Items an der selben Stelle befinden, wenn ein ganz bestimmtes Lösungsitem gefunden werden muss.

Abbildung 106: Eingekreist ist der Bereich, in dem sich das Zielitem befindet.

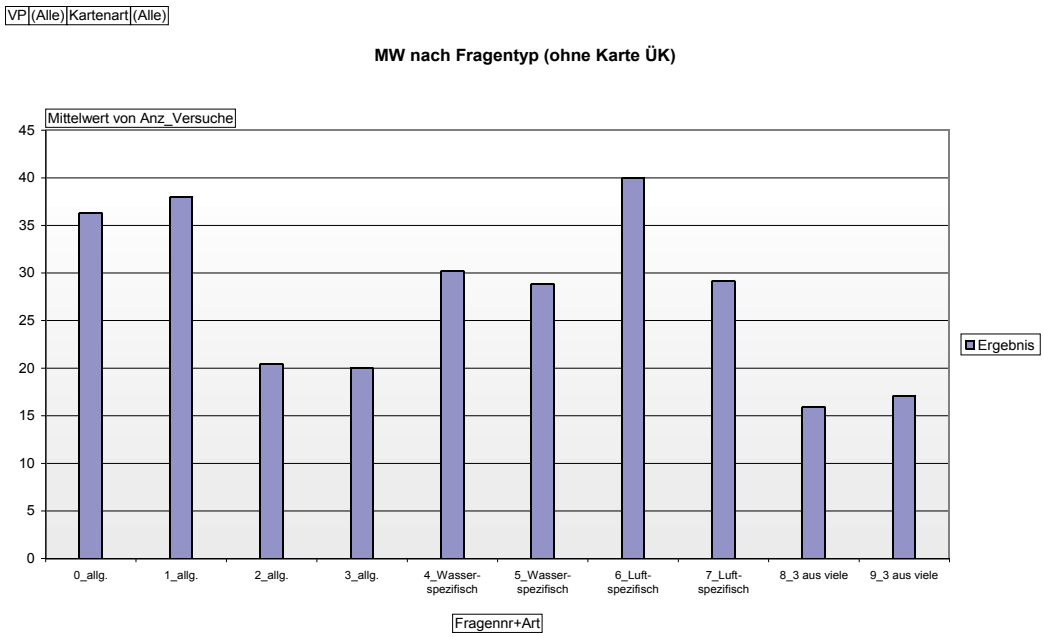


Abbildung 107: Mittelwerte der Fragen, zusammengefasst über alle Hofkarten

2. **Frage 0 mit Karte 3/8 (Luft, Wasser, Erde|Boden|Grund), MW 53.5, Abb. 108**

   Bei der Frage 0 (es geht um Erlebnisparks) sind zwei Items als Lösungen gültig, es musste aber nur eines gefunden werden. Der MW von Karte 3/8 ist mit 53.5 äusserst hoch. Auch hier spielt Verdeckung die Hauptrolle und erschwerend kommt das semantische Umfeld dazu. Beim oberen Item ist der unmittelbare Nachbarstext ein philosophischer Artikel und auch das nahe Cluster ist philosophisch – die VP wird also nicht dazu verleitet in der Nachbarschaft zu verweilen.

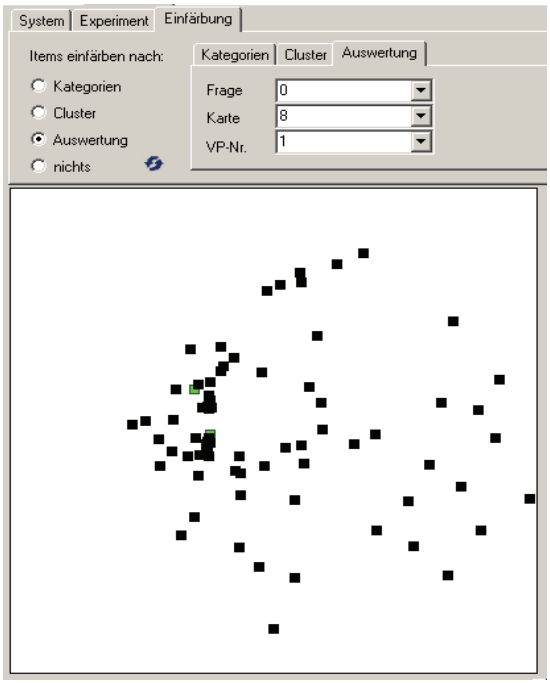
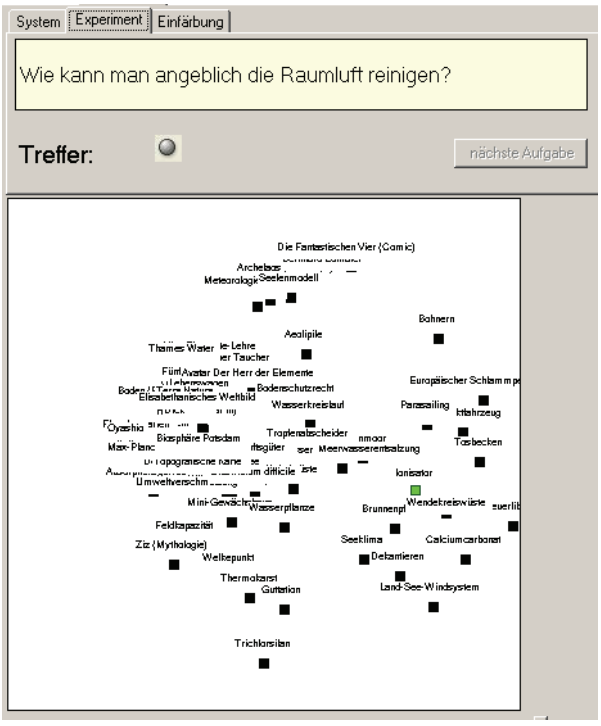


Abbildung 108: Die beiden Zielitems der Frage 0 sind verdeckt und befinden sich in einem themenfremden Bereich.

3. **Vergleich Fragen 7 und 4, mit Karte 3/8 (Luft, Wasser, Erde|Boden|Grund), Abb. 109**

   Die Karte 3/8 erweist sich bei der Frage 7 als nützlicher (MW 15.5, Anzahl Versuche: 1, 6, 11, 18, 19, 38) als die anderen Karten. Das Lösungsitem befindet sich in einer exponierten Lage. Warum



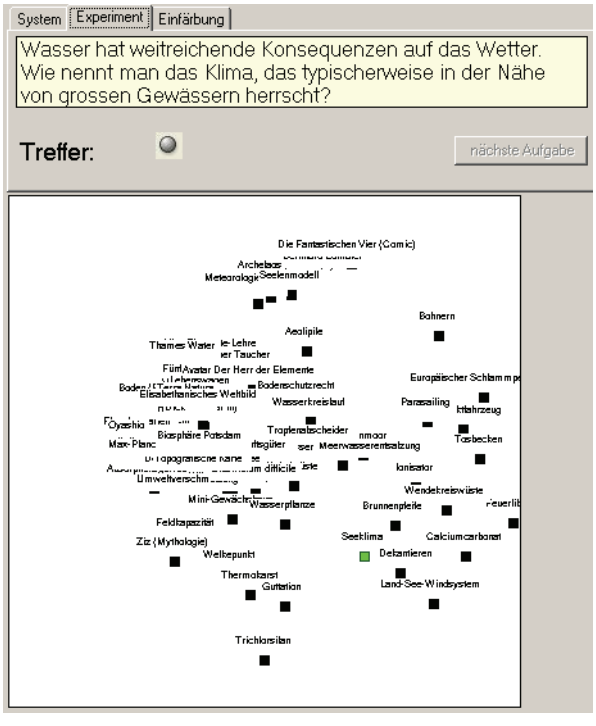


Abbildung 109: Karte 3/8 in Kombination mit Frage 7 (links) und 4 (rechts): Das Zielitem der Frage 7 wird drei Mal schneller gefunden als dasjenige der Frage 4, obwohl die Lage der Zielitems durchaus vergleichbar ist.

allerdings die Frage 4, deren Lösungsitem sich auch an einer exponierten Lage befindet, mit einem MW von 48.5 (Anzahl Versuche: 11, 15, 62, 62, 68, 73) drei Mal schlechter abschneidet, ist unklar.

4. **Frage 5 mit Karte 4/9 (ÜK), MW 55, Abb. 110**
   Die Karten, die auf dem ÜK beruhen, sind allgemein schlecht geclustert. Die VPn können nicht anhand der Strukturierung Themenbereiche ausmachen, sondern müssen die Themengrenzen durch Aufdecken ausfindig machen. Dementsprechend sind die MW bei den ÜK-Karten etwas höher.

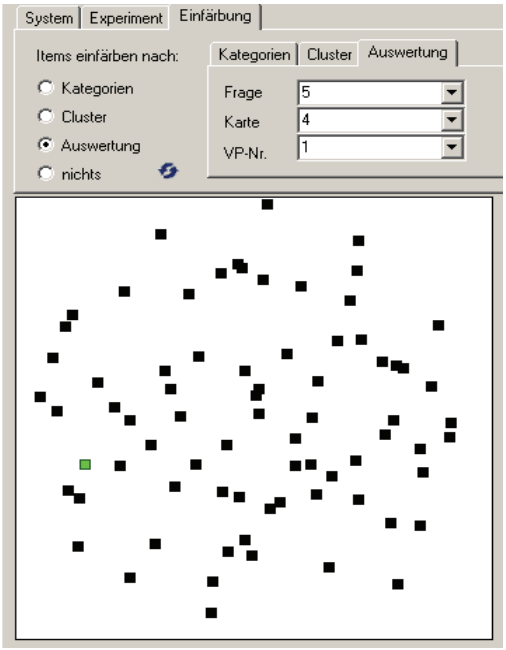


Abbildung 110: Unklare visuelle Strukturierung bei den ÜK-Karten, dementsprechend ist ihr MW etwas höher.

Besonders gute Konstellationen:

5. **Frage 2 mit Karte 2/7 (Erde|Boden|Grund), MW 9.5, Abb. 111**
   Die Frage zielt auf einen eindeutig philosophischen Text. In der Karte befinden sich rund 9 Texte, die mit Philosophie zu tun haben. In allen Karten clustern diese Texte ziemlich gut, so auch – und vor allem – in Karte 7. Das Zielitem befindet sich genau im Zentrum des Philosophie-Clusters. Bei der Karte 0 hingegen ist das Zielitem zwar auch im philosophischen Bereich, jedoch ist das Cluster weniger deutlich und das Item befindet sich weniger im geografischen Mittelpunkt. Lehre daraus: In einer optimalen Karte befindet sich das Zielitem unverdeckt im Zentrum eines passenden Clusters.

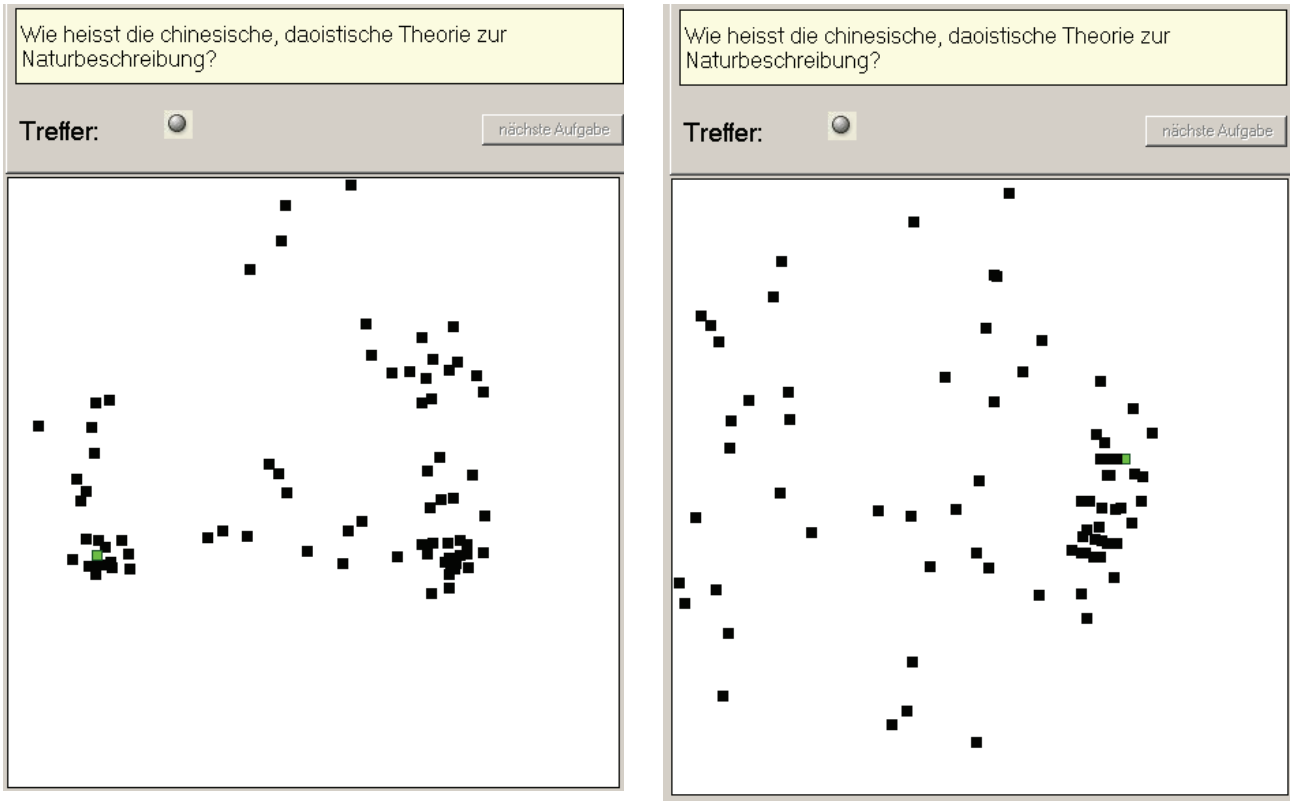


Abbildung 111: Das philosophische Zielitem befindet sich im Zentrum des Philosophie-Clusters und wird somit äusserst rasch gefunden.

Betrachten wir die Aufgabenkategorie «3 aus vielen» genauer:

6. **Frage 8 und 9 mit Karte 4/9 (ÜK), MW 7.5, Abb. 112**

   Die undeutliche Strukturierung der ÜK-Karten muss kein Nachteil sein, wie die Karten 4/9 beweisen. Sämtliche Lösungsitems liegen im selben Bereich. Ist einmal eines gefunden, kann die VP in dessen Nachbarschaft mit grosser Wahrscheinlichkeit weitere finden. Anders hingegen verhält es sich bei Frage 9: Die Antworttexte sind über die gesamte Karte verteilt und lassen sich schlecht «abklicken».

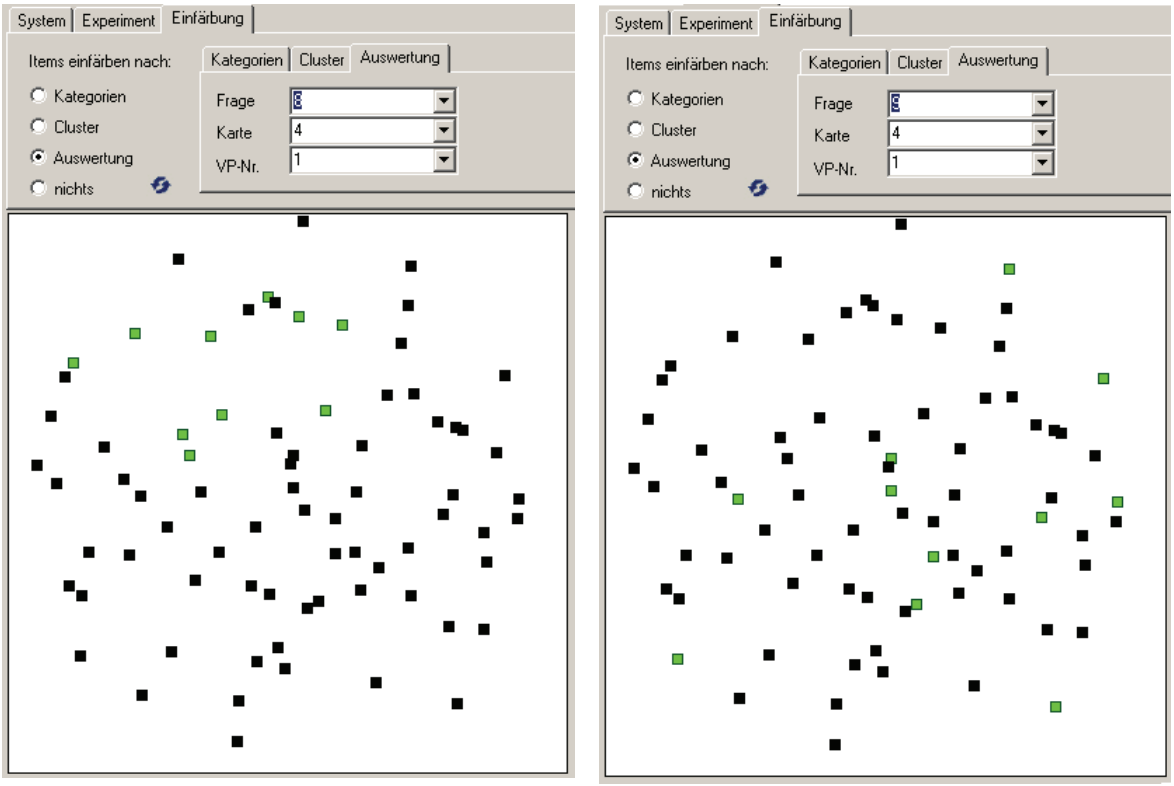


Abbildung 112: Links: Alle möglichen Lösungsitems der Frage 8 sind im selben Bereich. Die undeutliche Strukturierung der ÜK-Karte ist hier kein Nachteil. Die Lösungsitems der Frage 9 hingegen verteilen sich bei der ÜK-Karte über den gesamten Bereich und können so sehr schlecht gefunden werden.

7. Frage 8, Karten 0/5 (Luft) und 1/6 (Wasser), Abb. 113

Die Frage zielt auf das philosophische Cluster. Obwohl die beiden Karten und die Verteilung der Lösungsitems auf den ersten Blick ähnlich aussehen, benötigen die VPn bei der Karte «Luft» doppelt so lange, als bei der Karte «Wasser». Die Erklärung ist offen. Ein Hinweis liefert vielleicht die Prokrustes-Transformation: Ein Average Loss von 0.63 ist recht hoch; die Karten sehen äusserlich betrachtet ähnlich aus, sind von der Struktur her jedoch unterschiedlich. Im nächsten Abschnitt wird mehr auf diese Kartenkonsistenz eingegangen.

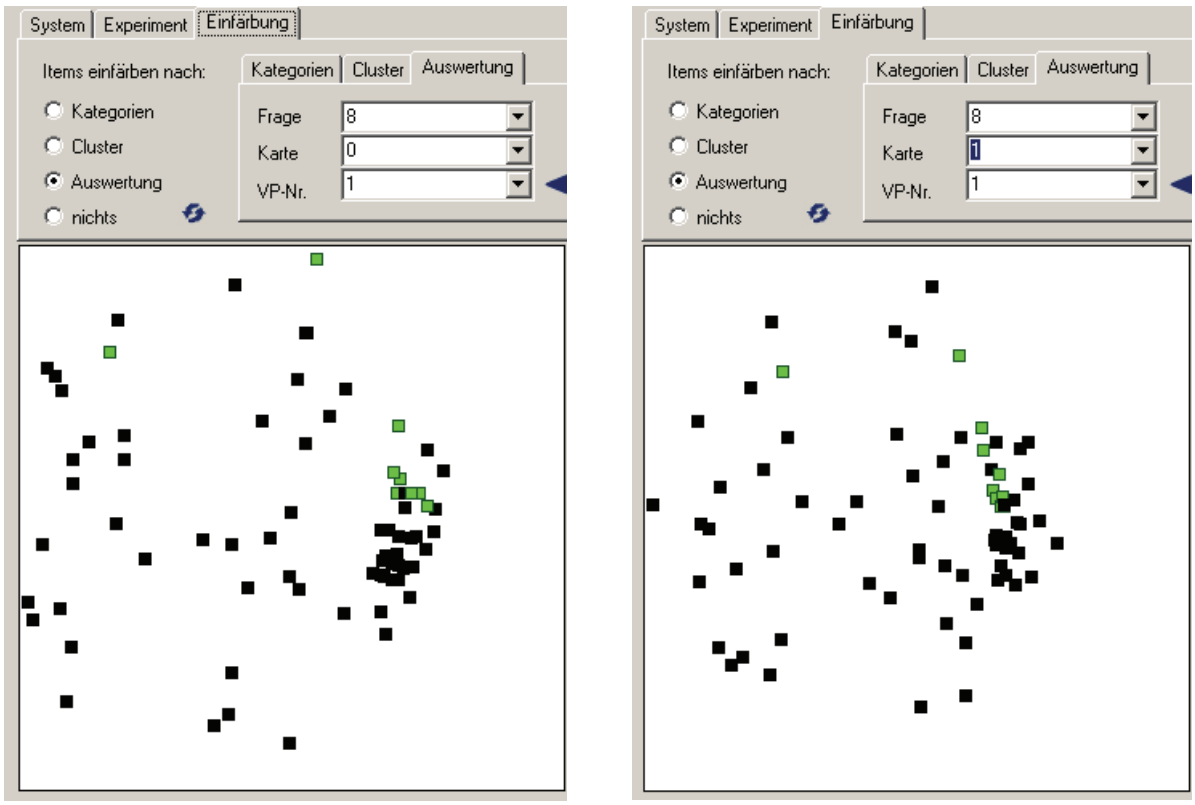


Abbildung 113:Wieso brauchten die VPn für die linke Karte doppelt solange (MW 22), um drei der Zielitems zu finden, als auf der rechten Karte (MW 11.67)? Die linke Karte wurde nach dem Stichwort «Luft» behoft, die rechte nach «Wasser». Die Frage war jedoch eine philosophische.

24.4.2 Kartenkonsistenz

Die Karten unterscheiden sich untereinander stark in der Anordnung der einzelnen Items. Zwar sind in allen Karten die philosophischen Texte geclustert, ebenso sind die naturwissenschaftlichen Text nahe beieinander, jedoch variieren die Positionen der umliegenden Texte sehr stark. Abbildung 114 zeigt die verschiedenen Prokrustes-Transformationen aller HM-Karten (die ÜK-Karten weichen sowieso noch stärker von dieser Grundkonfiguration ab). Die Verschiebungen ausserhalb der Philosophie- und Wissenschaftscluster sind beträchtlich.

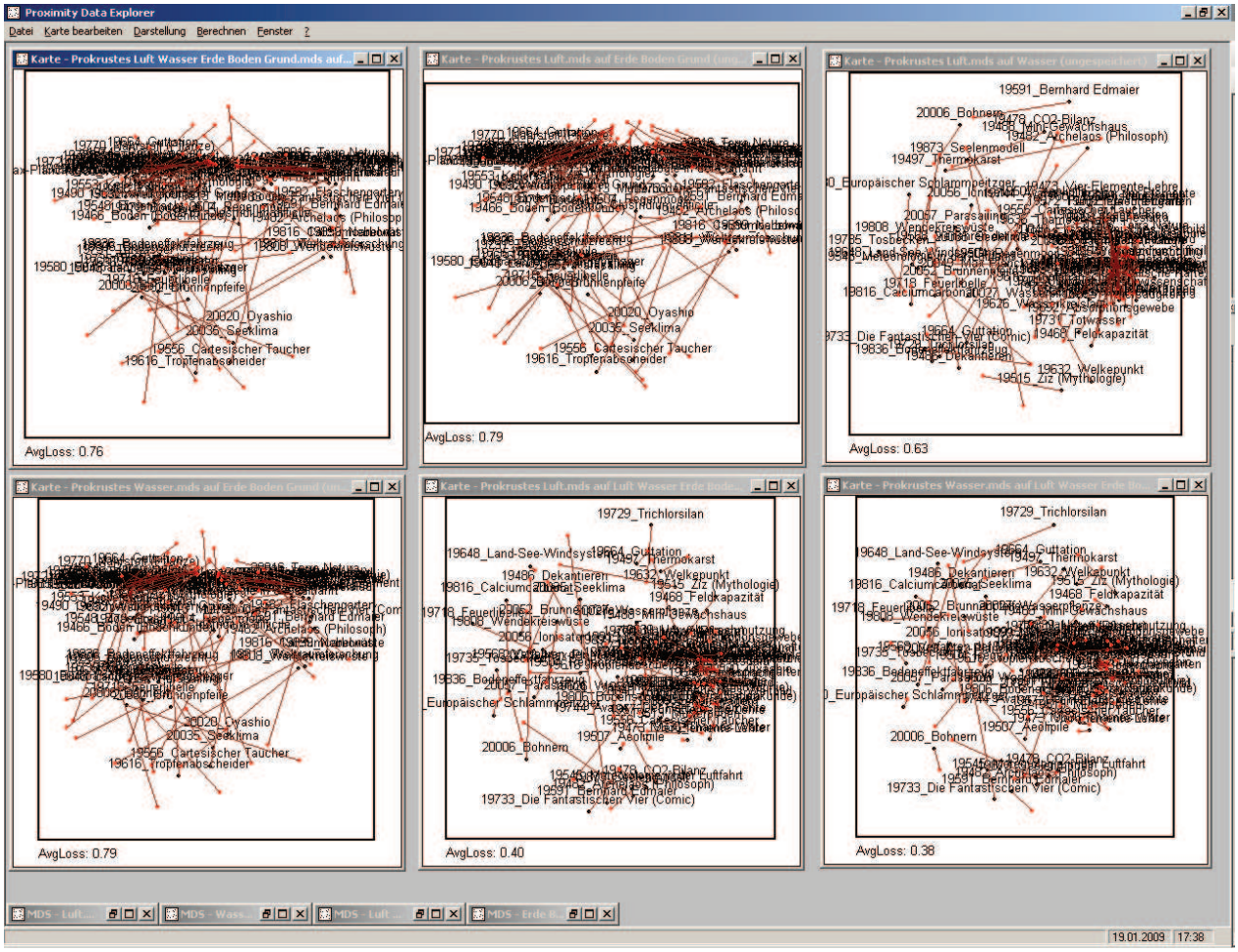


Abbildung 114: Prokrustes-Transformationen aller Karten. Die unterschiedlichen Konfigurationen sind deutlich sichtbar. Nicht abgebildet sind die ÜK-Karten, die jedoch noch stärker von diesen Konfigurationen abweichen.

24.5 Resultate: Personenfaktoren

Die Fragen lassen sich in die zwei Kategorien «Einzelaufgaben» und «3 aus vielen» aufteilen. Zur ersten Kategorie gehören nur diejenigen Fragen, bei denen es genau ein Lösungsitem gab, zur zweiten Kategorie gehören die beiden Fragen der gleichnamigen Kategorie (die allgemeine Frage, bei der es zwei Lösungsitem gab, wird in diesem Kapitel nicht ausgewertet).

Bei der Kategorie «Einzelaufgaben» wäre bei 76 Items und bei komplett zufälliger Ziehung ein MW von 38 zu erwarten, tatsächlich liegt er aber bei 29.5 (ohne Karte ÜK). Die Werte verteilen sich nicht um diesen MW, sondern um den Wert 10 (s. Abb. 115). Man könnte daraus interpretieren, dass einige VPn die Aufgaben gut lösen, während einige wenige VPn äusserst schlecht abschneiden und den MW somit herauf drücken. Schaut man sich aber Klickraten der Einzelpersonen an, ist es eher so, dass eine Mehrheit der VPn

die meisten der sieben Einzelfragen relativ bald findet, für ein paar wenige Fragen jedoch sehr lange braucht. Die sehr hohe Standardabweichung von 20.87 unterstreicht dies.

Zusätzlich zum Experiment generierten wird ein Zufallsset. Es galten die gleichen Bedingungen wie beim Experiment, nur wurden die Texte in der Simulation zufällig abgeklickt.

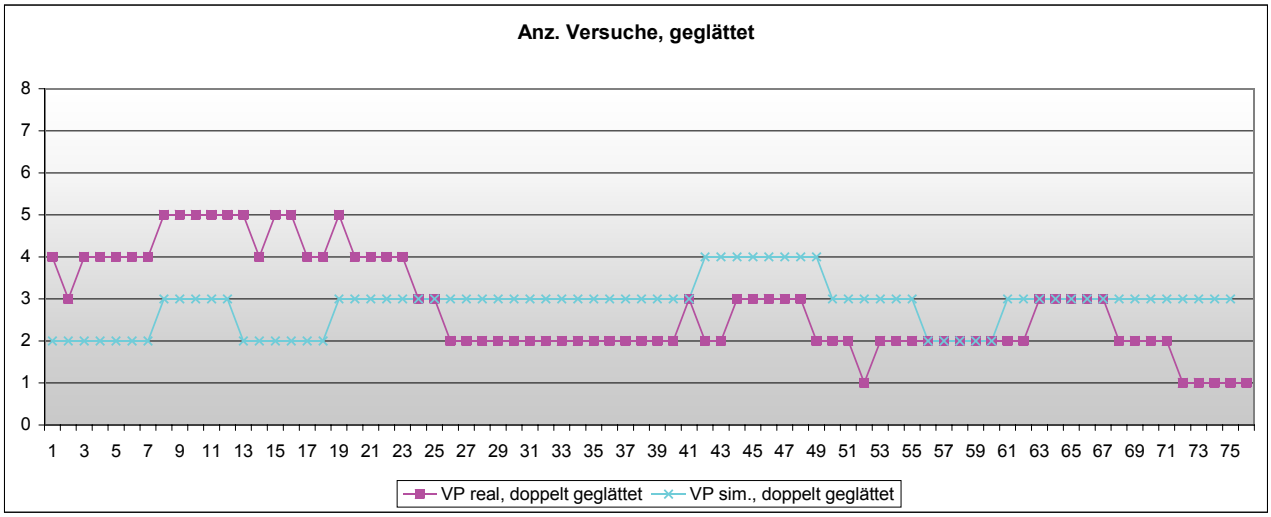


Abbildung 115: Es wird nur die Aufgabenkategorie «Einzelaufgaben» betrachtet. Die benötigte Anzahl Versuche ist bei den realen VPn nicht um den MW von 31 normalverteilt, sondern hat seinen Höhepunkt um den Wert 10. Bei den simulierten VPn ist die erwartete Gleichverteilung bei ca. 2.8 grob ersichtlich.

Die Verteilung dieses Zufallssets ist erwartungsgemäss näherungsweise um den Wert 3 gleichverteilt.

Der MW des Zufallssets liegt für die Kategorie «Einzelaufgaben» bei 40.06 und ist somit hochsignifikant höher als bei den realen VPn. Die Standardabweichung ist mit 21.08 nur unwesentlich höher (s. Abb. 116).

Erstaunlicheres tritt bei der Fragenkategorie «3 aus vielen» zu Tage: der MW des Zufallssets (20.13) ist nicht signifikant höher als bei den realen VPn und die Standardabweichung ist praktisch gleich (real: 9.44, simuliert: 9.47)! Das von den VPn positiv erlebte Ergebnis dieser Fragenkategorie ist also der zufälligen Wahrscheinlichkeit gleich.

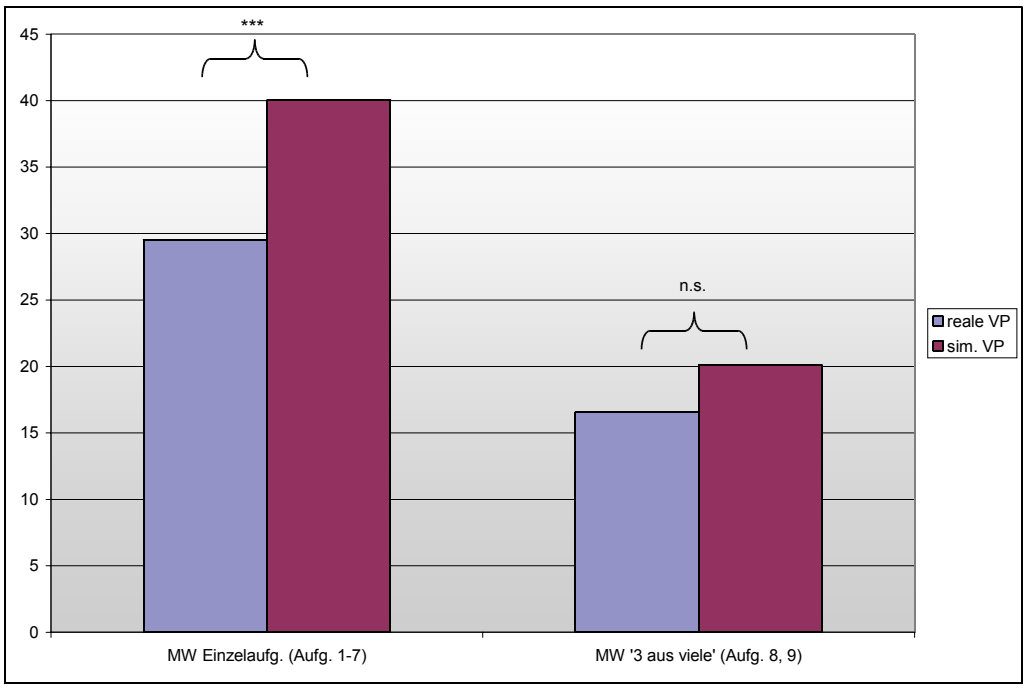


Abbildung 116: Deutliche Unterschiede im MW zwischen den realen und simulierten VPn bei der Fragenkategorie «Einzelaufgaben», jedoch nicht bei der Kategorie «3 aus vielen»

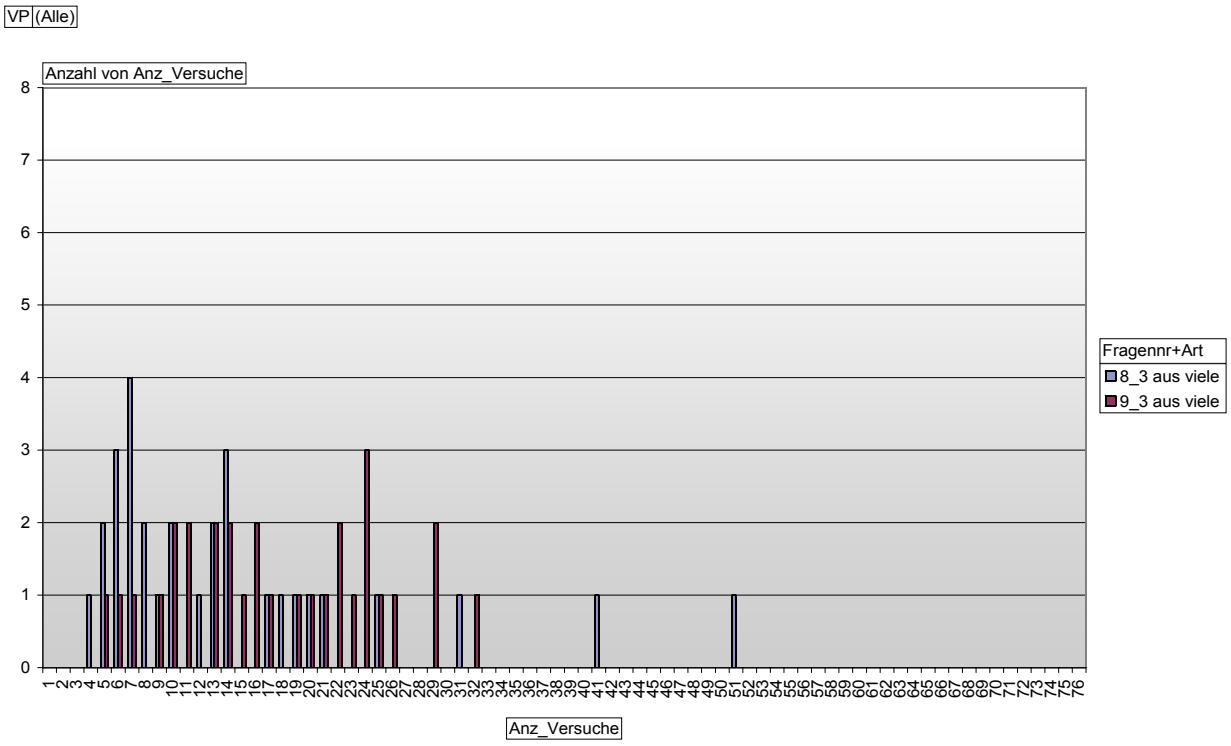


Abbildung 117: Auch bei der Aufgabenkategorie «3 aus vielen» ist die Verteilung der Anzahl Versuche deutlich linkslastig.

Die Frage 8 aus der Aufgabenkategorie «3 aus vielen» ist deutlich linkslastig (s. Abb. 117); die meisten VPn haben das Lösungscluster rasch gefunden und dann durchgeklickt. Die Frage 9 hingegen ist nur unwesentlich besser als die Zufallserwartung von 22.

Die Hypothese 1 (die Strukturierung hilft bei der Kartensuche) wurde bei den Einzelaufgaben klar bestätigt, bei den «3 aus vielen»-Aufgaben nicht.

24.5.1 Einzelbeispiel

Obwohl wir in diesem Experiment die Klickpfade (in welcher Reihenfolge die Items angeklickt wurden) der VPn nicht aufgezeichnet hatten, lässt sich anhand der Experimentsprotokollierung wenigstens an einem Beispiel zeigen, was für Überlegungen sich eine aufmerksame VP machen kann.

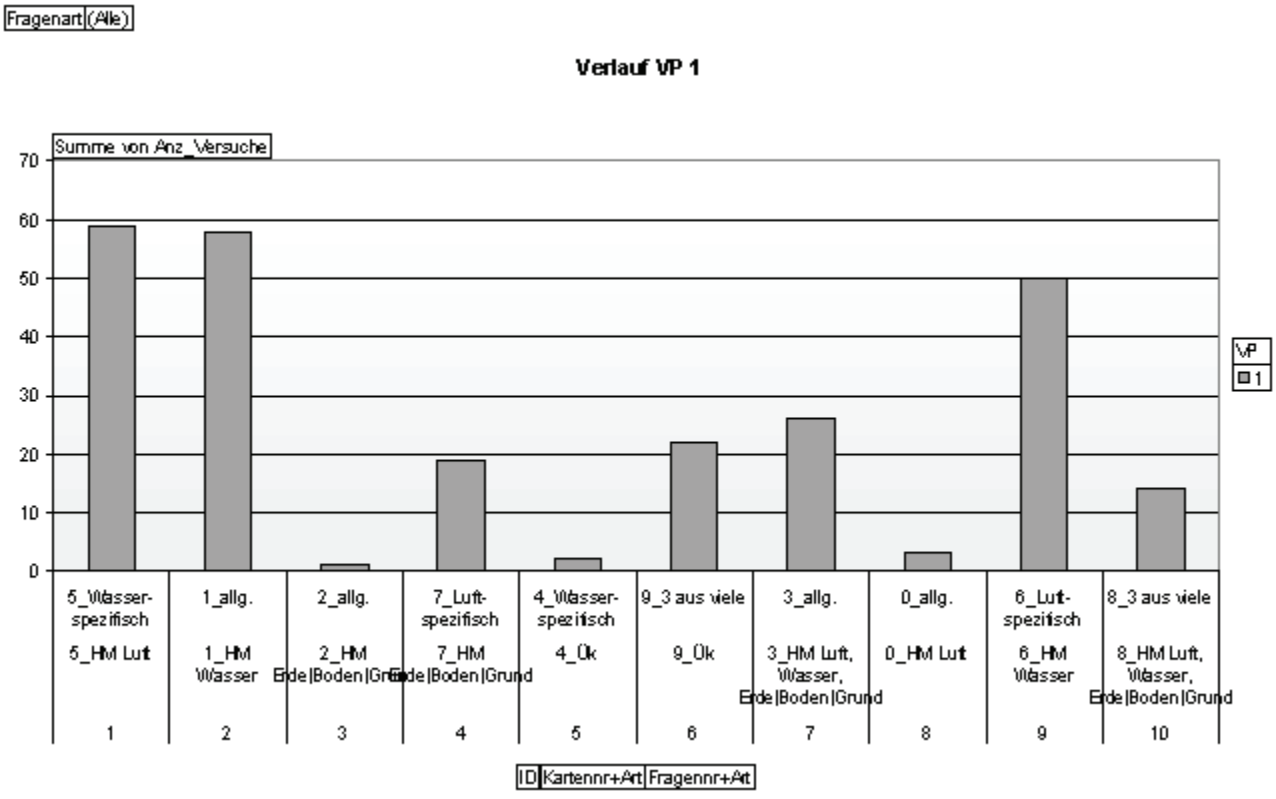


Abbildung 118:Verlauf der benötigten Klicks der VP 1.

Die VP wirkte bei den ersten zwei Aufgaben verzweifelt. Als sie mit der zweiten fertig war, hatte sie aber das Gefühl, die Logik der Karten begriffen zu haben. Die dritte Frage lautete: «Wie heisst die

chinesische, daoistische Theorie zur Naturbeschreibung?». Die VP erinnerte sich, dass es in den Karten zuvor ein philosophisches Cluster gegeben hatte und zielte in der neuen Karte mittig auf das engste Cluster: Treffer beim ersten Klick!

Die Frage zur 8. Aufgabe lautete: «Als Höhepunkt möchtest du mit der Klasse eine Reise unternehmen und einen Erlebnispark besuchen, der sich mit deiner Thematik beschäftigt. Finde solch einen Park!». Die VP runzelte die Stirn und meinte, dieses Item klinge aussergewöhnlich und müsse wohl etwas abseits liegen. Prompt erwischte sie dieses schwierige Zielitem beim 3. Klick.

24.6 Diskussion

Zusammenfassend lässt sich festhalten, dass weniger die Art der Kartenberechnung, also ob HM oder ÜK, wichtig ist, sondern die konkrete Lage des Zielitems. Erstaunlicherweise spielt es auch keine grosse Rolle, nach welchem Stichwort die zu Grunde liegenden Texte behoft wurden. Die Luft-Karte beispielsweise führt bei der Luft-Frage nicht zu einem besseren Resultat als die Wasser-Karte. Als bedeutsamer erwies sich der thematische Inhalt der Wikipedia-Artikel: Die philosophischen Texte clusterten bei allen Karten sehr deutlich, ebenfalls waren die naturwissenschaftlichen Umweltartikel enger benachbart. Eine deutliche Clusterung zieht die Aufmerksamkeit der VP auf sich und verleitet zum Reinklicken.

Ist anzunehmen, dass ein ganz bestimmter Zieltext gefunden werden soll, muss sichergestellt werden, dass sich die Items in der Karte nicht verdecken. Idealerweise befindet sich das Zielitem im Zentrum seines thematischen Clusters.

24.7 Ausblick

Bei einem Folgeexperiment müssen unbedingt die «Klickpfade» festgehalten werden: In welcher Reihenfolge klickt die VP die Items ab; was sind die Suchstrategien? Es liesse sich dann genauer untersuchen, durch welche Konfigurationen sich VPn verwirren und vom Zielitem ablenken lassen. Des Weiteren müssen die VPn durch einfachere Aufgaben motiviert werden. Es zeigte sich, dass viele VPn schnell das Gefühl hatten versagt zu haben und frustriert waren. Weitere Idee: Um die gewonnenen Erkenntnisse umzusetzen und zu testen könnte man eine «optimale» Karte konstruieren, in der das Zielitem am ehesten gefunden wird. Diese Karte wäre der VP am Schluss des Durchganges zu präsentieren, damit sie nicht geprimed würde.

25 Homonyme: Explorationsexperiment Bach/Golf

25.1 Überblick

Die Hofmethode erlaubt aufgrund ihrer Kontextabhängigkeit die Disambiguierung von Homonymen. In diesem Explorationsexperiment werden die verschiedenen Bedeutungen der Homonyme Bach und Golf in Wikipedia-Artikeln durch die HM berechnet und in semantischen Karten sichtbar gemacht.

25.2 Einleitung

Wie im ersten Teil dieser Arbeit beschrieben, bezieht die HM den Kontext von Stichwörtern in die semantische Ähnlichkeitsberechnung mit ein. Dadurch weist sie Parallelen zu unserem Hirn auf, das ausgesprochen kontextabhängig arbeitet. Erst dadurch ist es uns möglich in einer Konversation unverstandene Wörter mental zu ergänzen. Auch fällt es uns leicht, die korrekte semantische Auslegung eines Homonyms zu erraten. Diese Homonymdisambiguierung sollte auch für die Hofmethode ein leichtes zu vollziehen sein: Identische Wortformen mit unterschiedlichen Bedeutungen müssen sich durch ihren unterschiedlichen Kontext auszeichnen. Wir machten die Probe aufs Exempel mit realen Texten und verglichen zwei Verfahren zur Keyword-Bestimmung anhand zweier Homonyme.

25.3 Vorgehen

Als Homonyme suchten wir «Bach» und «Golf» aus. Mit «Bach» kann ein Gewässer oder der Komponist gemeint sein, mit «Golf» die Automarke, der Sport oder der Meerbusen. Als Datengrundlage dient – wie schon im Kapitel 24 (Wikipedia-Experiment) – die Wikipedia[31], deren Inhalt auf den lokalen Computer gespiegelt und mit der Open-Source-Suchmaschine Lucene[32] indiziert wurde. Die Texte wurden nach den Stichwörtern «Bach» und «Golf» durchsucht und die jeweils ersten 80, beziehungsweise 60 Texte der retournierten Lucene-Liste wurden in die Datenbank des SemanticMappers aufgenommen. Den Texten wurde manuell eine Kategorie zugeteilt, die zwischen den unterschiedlichen Wortbedeutungen

31 http://www.wikipedia.org
32 http://lucene.apache.org/

differenzierte. Bei den Berechnungen wurde diese Kategorie selbstverständlich nicht berücksichtigt, jedoch konnten so die Karten im Nachhinein eingefärbt werden, um die Augenscheinvalidität beurteilen zu können.

Die TargetWords beider Textsammlungen wurden mit dem KWII-Verfahren (s. Kap. 15; KeywordII-Analyse) und dem Tagcloud-Verfahren (s. Kap. 16; Tagcloud-Verfahren von Semager) bestimmt und behoft. Die beiden resultierenden Karten beider Textsammlungen wurden anschliessend miteinander verglichen.

25.4 Resultate

25.4.1 Bach-Texte

Beide Verfahren produzieren sehr ähnliche Karten: Die Karten sind in drei Bereiche aufgeteilt. In einem Bereich sind die Gewässer-Texte, in einem anderen alle Texte, die mit dem Komponisten Johann Sebastian

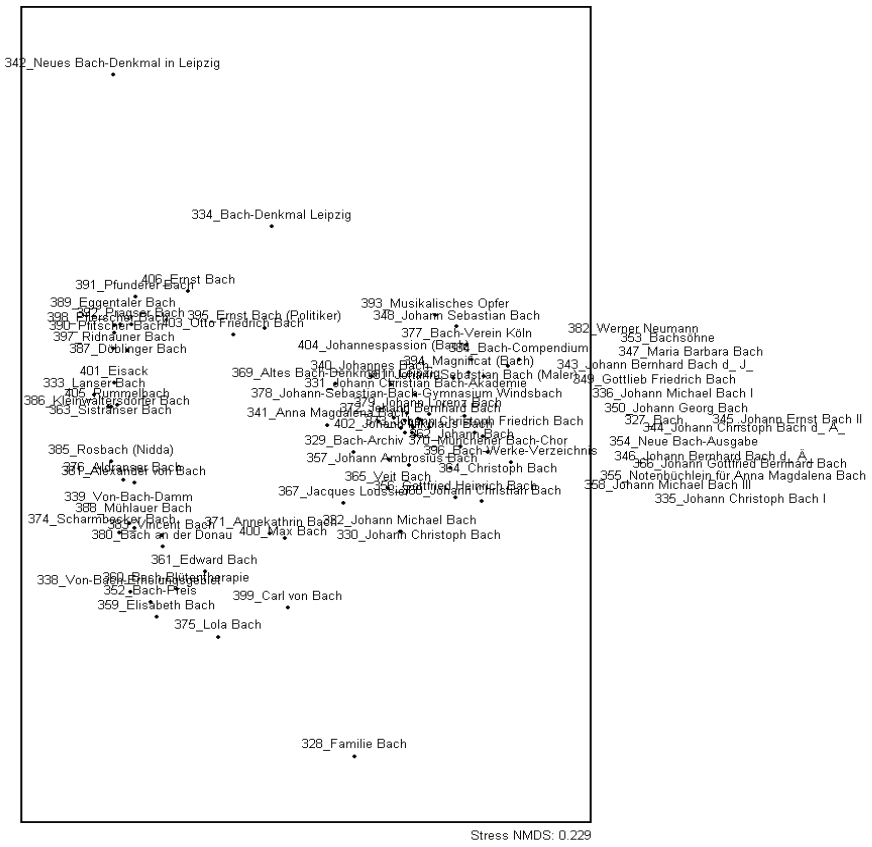


Abbildung 119: Die Texte dieser Karte wurden mit dem KWII-Verfahren behoft. Links sind die Gewässer-Texte, rechts alle Texte, die mit dem Komponisten zu tun haben, dazwischen Personen, die zwar Bach heissen, aber nicht mit J.S. Bach verwandt sind.

Bach oder dessen Familienangehörigen zu tun haben (auch Denkmäler), dazwischen sind vor allem übrige Personen, deren Namen «Bach» lautet. In Abb. 119 ist die Karte dargestellt, die mit dem KWII-Verfahren erstellt wurde. Die Items sind mit den Wikipedia-Artikeltiteln beschriftet. Der rechte Bereich (derjenige mit dem Komponisten Bach) ist sehr dicht; damit die Labels trotzdem lesbar sind, wurden sie manuell nach rechts und auseinander gezogen. Ausschlaggebend für die Betrachtung der Struktur sind die Text-Punkte, nicht die Label.

In Abbildung 120 ist dieselbe Karte abgebildet, jedoch sind die Texte nach der zugewiesenen Kategorie eingefärbt. Innerhalb des Komponisten-Bach-Clusters rechts strukturieren sich die Texte sogar tendenziell nach der Person, beziehungsweise dessen Angehörigen (blau, unterer Bereich) und Artefakten, die mit dem Komponisten zu tun haben (türkis, oberer Bereich). Die Trennung ist allerdings nicht scharf.

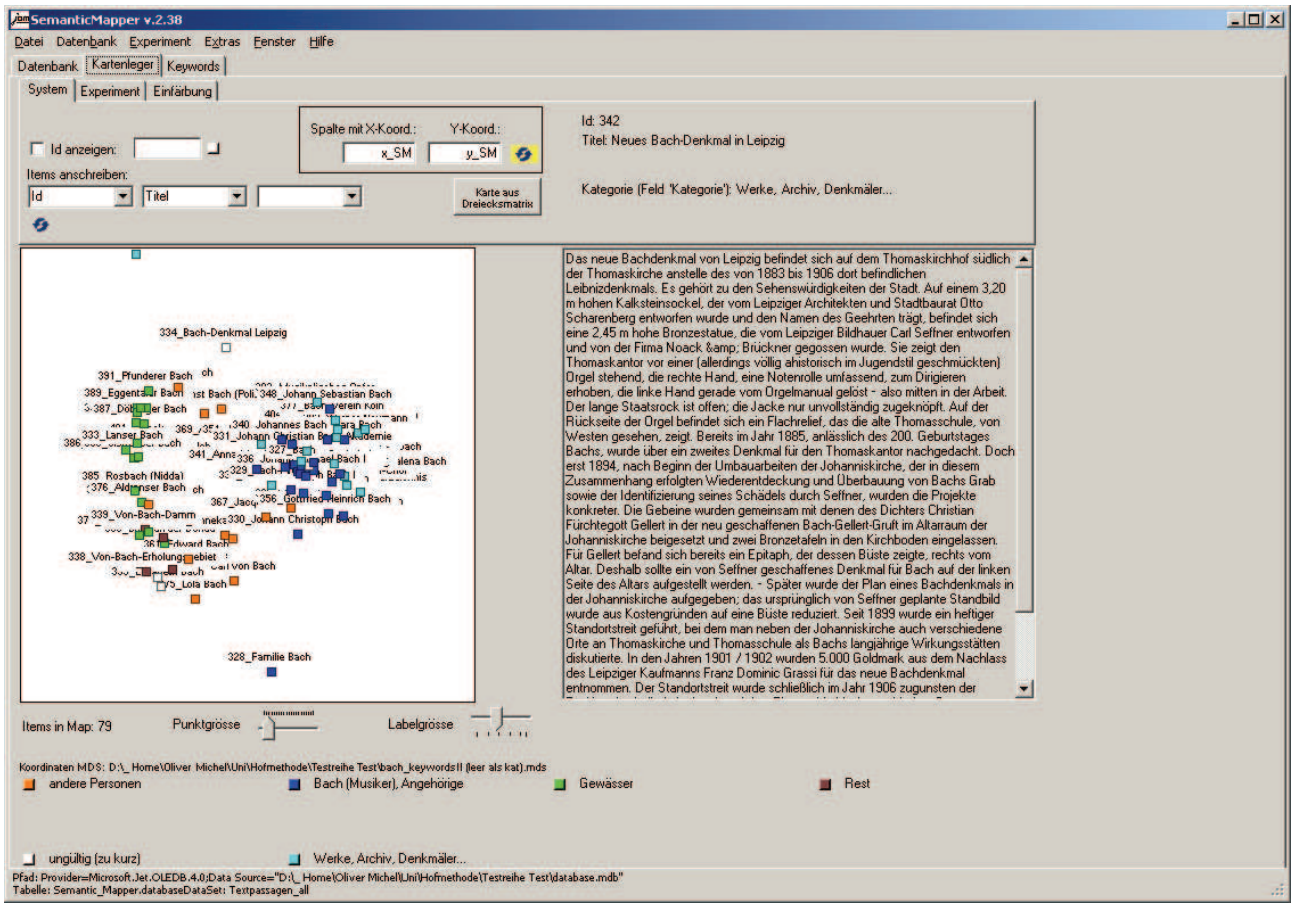


Abbildung 120: Die eingefärbten Texte der KWII-Karte - deutlich sind die verschiedenen Bereiche erkennbar. Die erklärende Legende ist im unteren Teil der Abbildung sichtbar.

Durch den hohen Anteil an Komponist-Bach-Texten wird die Kartenstruktur durch diese Texte bestimmt. Die anderen Texte werden so gut es geht in diese Struktur eingepasst. Dies erklärt, warum der Bereich der Gewässer- und der Übrigen-Personen-Texte so langgestreckt, beziehungsweise aufgeteilt ist.

Zwei Texte liegen im peripheren Bereich («342_Neues Bach-Denkmal in Leipzig» ganz oben und «328_Familie Bach» ganz unten). Diese Platzierung ist verwirrend und anhand der Textinhalte nicht nachvollziehbar.

Die Karte basierend auf den TargetWords aus Tagcloud sieht grundsätzlich sehr ähnlich aus (Abb. 121), was auch die Prokrustes-Transformation belegt (Abb. 122). Der Ausreisser «328_Familie Bach» bleibt bestehen, jedoch nicht «342_Neues Bach-Denkmal in Leipzig», dafür gibt es links zwei neue Ausreisser «387_Döblinger Bach» und «388_Mühlauer Bach».

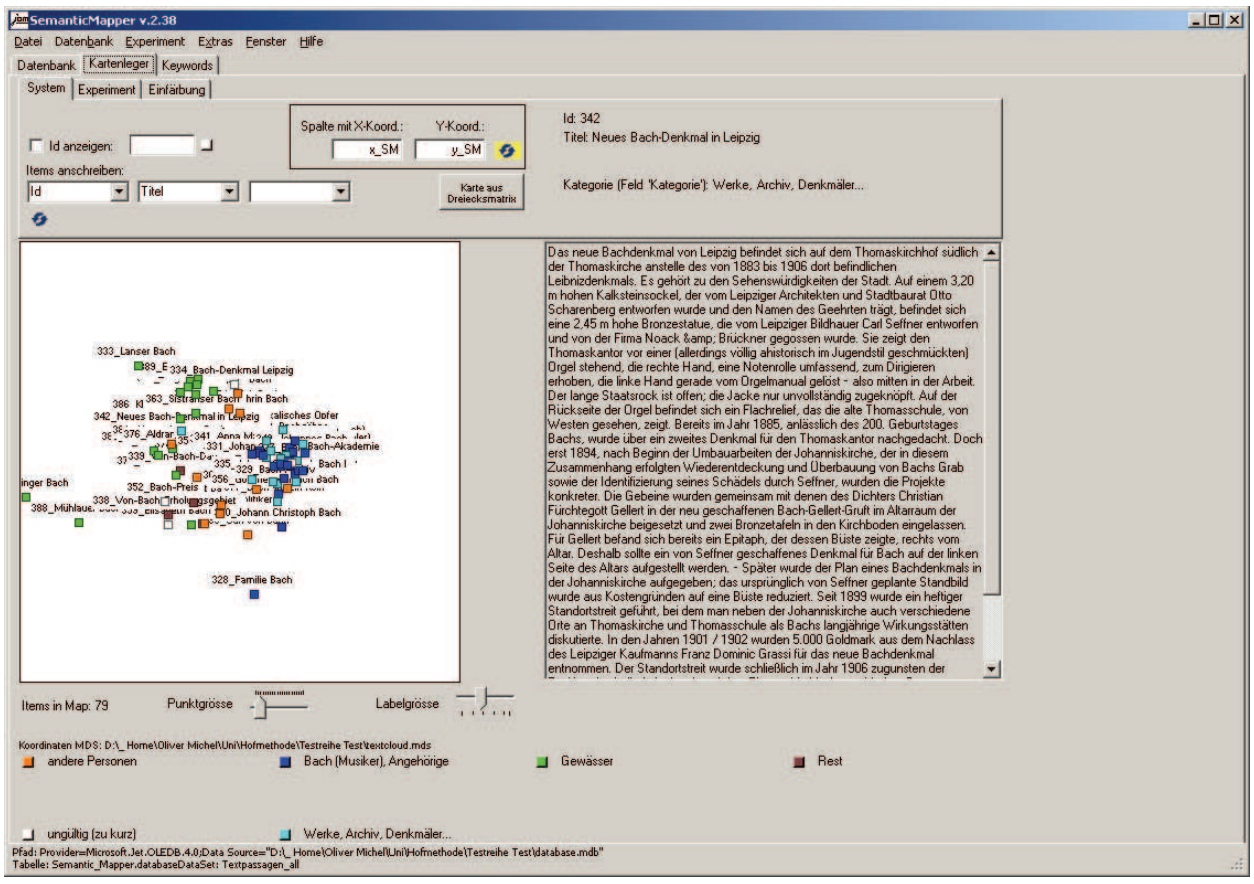


Abbildung 121: Die Karte basierend auf dem Tagcloud-Verfahren strukturiert sich sehr ähnlich wie die KWII-Karte.

Abbildung 122: Die Prokrustes-Transformation der KWII-Karte (schwarz) mit der Tagcloud-Karte (rot) zeigt die ähnliche Struktur: Die groben Bereiche sind ziemlich konsistent; es gibt nur sehr wenige Texte, die in semantisch anderen Bereichen liegen.

25.4.2 Golf-Texte

Auch bei den Golf-Texten produzieren das KWII- und das Tagcloud-Verfahren ähnliche Karten. Bei beiden fällt die abgelegene Position eines einzelnen Textes auf, während die anderen Texte eine recht kompakte Struktur bilden (am Beispiel der KWII-Karte in Abb. 123 gezeigt).

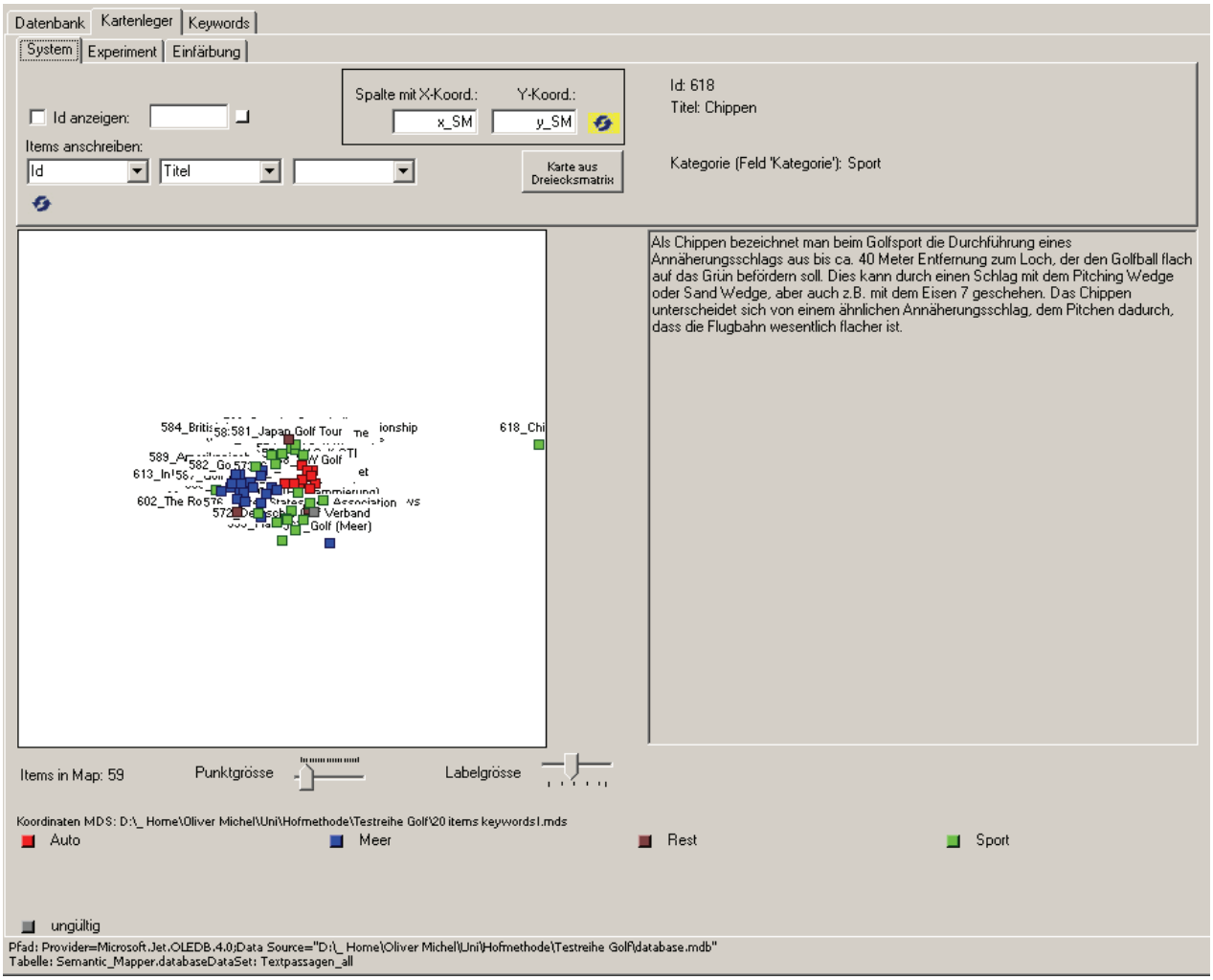


Abbildung 123: Abgesonderte Position des Textes «Chippen» (hier gezeigt anhand der KWII-Karte).

Der Titel des Aussenseiters lautet «Chippen» und beschreibt eine Schlagtechnik im Golfsport. Der Text ist sehr kurz und beide Verfahren finden nur ungünstige Zielwörter (s. Abb. 124).

Lässt man diesen Text weg, resultieren aussagekräftigere Karten (s. Abb. 125). Die drei semantischen Bereiche von «Golf» sind klar erkennbar. Zwar ist der Sportbereich zweigeteilt, jedoch bilden die fahrzeugverwandten Texte und die Meerestexte eigene Bereiche. Die linke Hälfte der Golfsporttexte besteht fast ausschliesslich aus Beschreibungen von Golfclubs. Das einzige Nicht-Golfclub-Item ist «599_Platzerlaubnis». Item «609_Micro Golf» liegt bei beiden Karten in den Meerestexten. Bei «Micro Golf» handelt es sich um eine Miniaturform von Golf, ähnlich wie Tischfussball. Es ist somit richtig, dass es ausserhalb des Golfsport-Bereiches liegt, allerdings spricht inhaltlich nichts dafür es bei den Meerestexten zu platzieren.

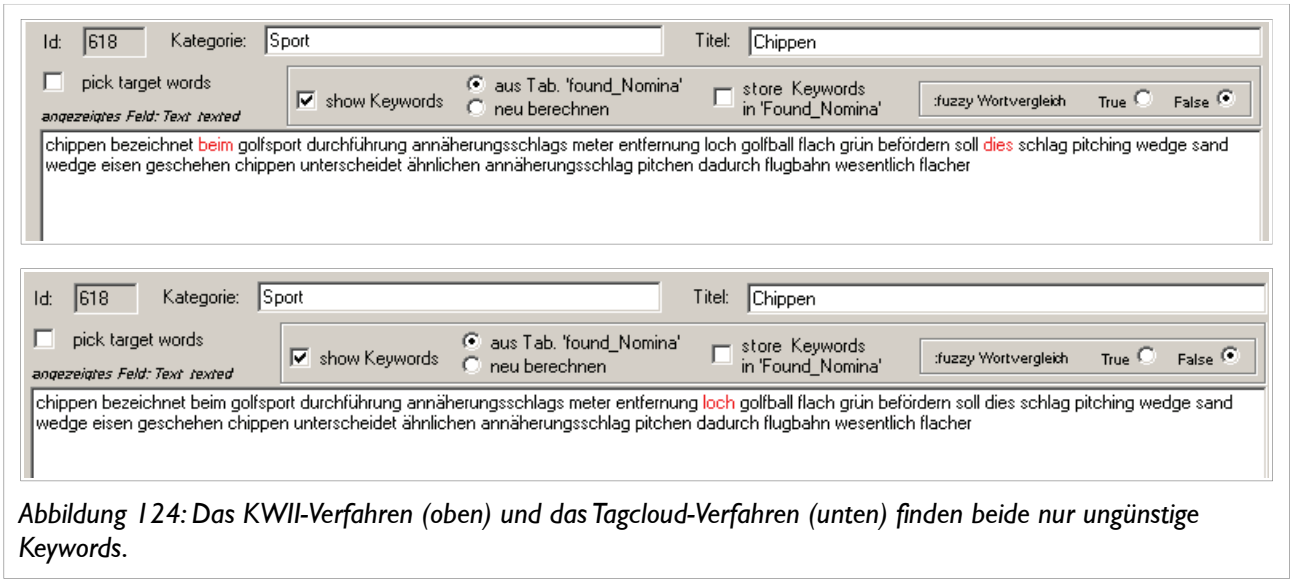


Abbildung 124: Das KWII-Verfahren (oben) und das Tagcloud-Verfahren (unten) finden beide nur ungünstige Keywords.

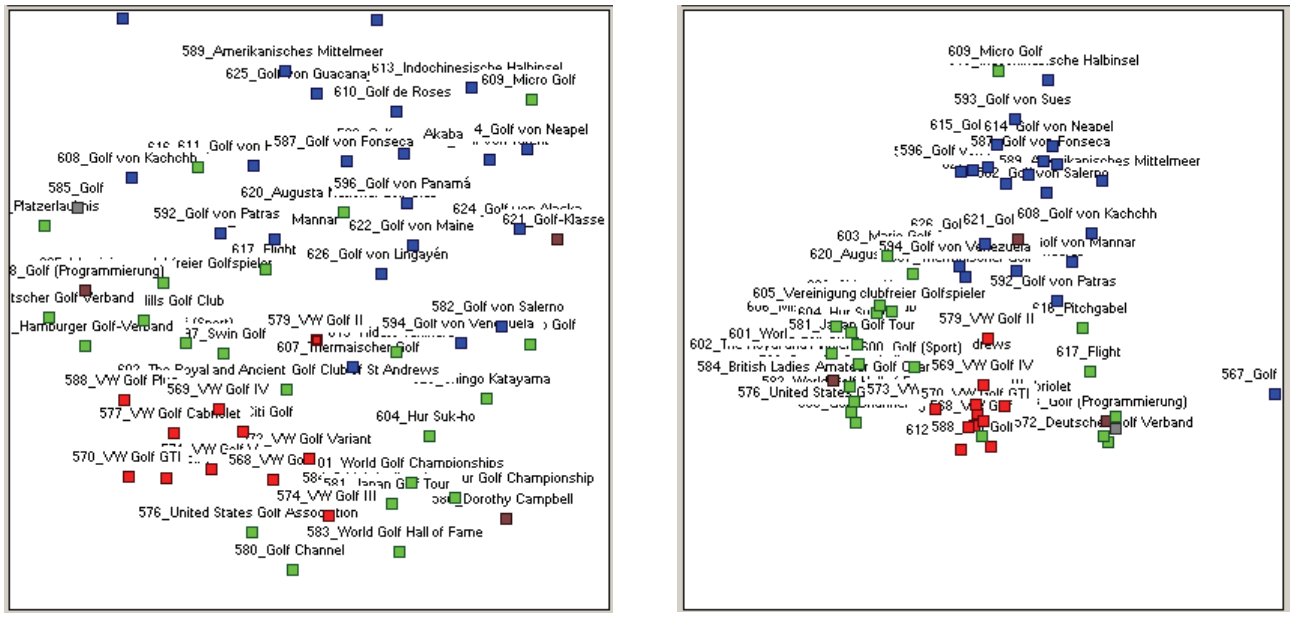


Abbildung 125: Die Fahrzeugtexte (rot) sind beisammen, ebenso die Meertexte (blau). Die Sporttexte (grün) sind zweigeteilt. Die Bereiche sind bei der KWII-Karte (links) weniger deutlich, als bei der Tagcloud-Karte (rechts).

Schliesslich zeigt Abbildung 126 die Prokrustes-Transformation der beiden KWII- und Tagcloud-verfahren. Auch mit den Golf-Texten produzieren die beiden Verfahren ähnliche Karten, wenn auch die Tagcloud-Karte deutlichere abgegrenzte Bereiche formt.

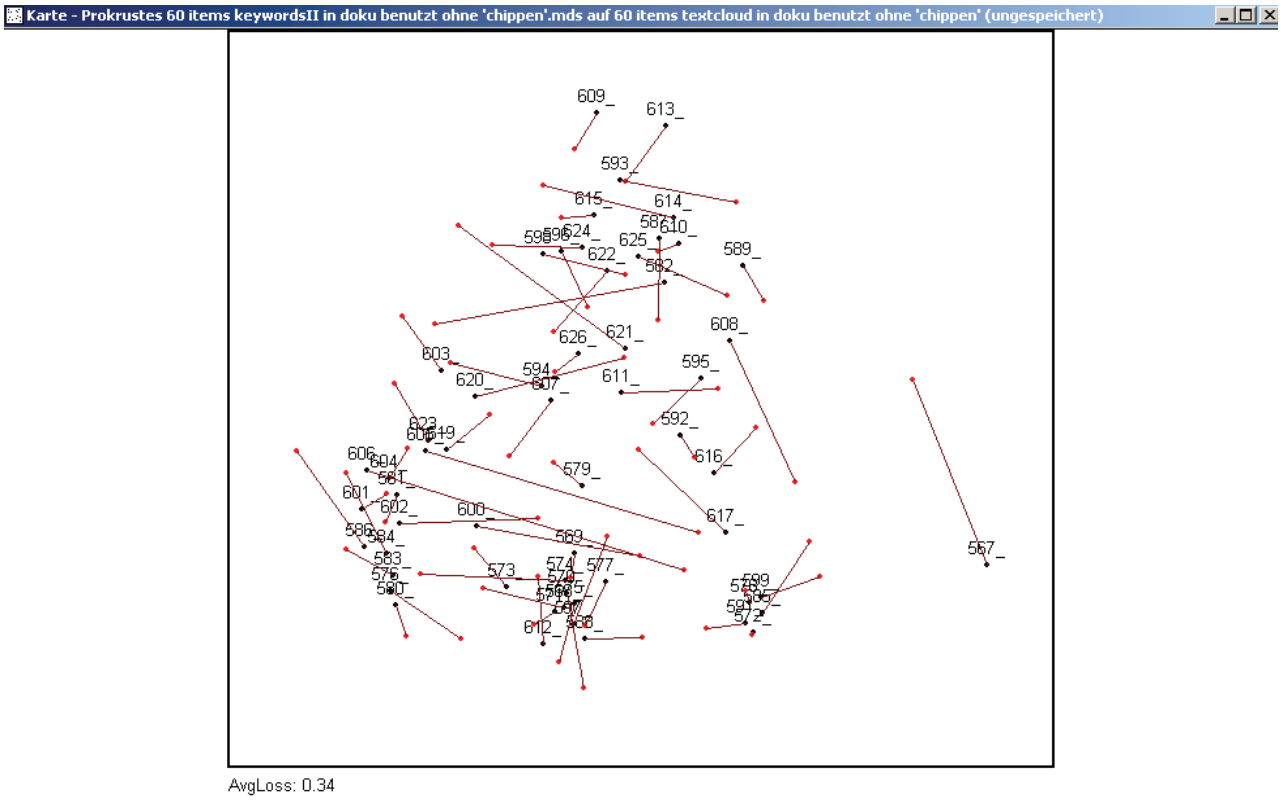


Abbildung 126: Die Prokrustes-Transformation der KWII- (rot) und Tagcloudverfahren (schwarz) zeigt deren hohe Ähnlichkeit.

25.5 Diskussion

Die Hofmethode erkennt in diesem Experiment die verschiedenen Bedeutungen der Homonyme «Bach» und «Golf» in sehr guter Weise. In den resultierenden semantischen Karten sind die Bereiche deutlich erkennbar, je nach verwendetem Keywordverfahren (KWII oder Tagcloud) sogar annähern geclustert (Tagcloud).

Unser psychologischer Ansatz, Wortbedeutung anhand der Höfe zu berechnen, erweist sich gerade im Umgang mit Homonymen als überaus fruchtbar. Es ist einfach plausibler, Wörter anhand ihres Kontextes zu evaluieren, anstatt simple Wortformen zu vergleichen – was gleich aussieht, muss nicht Gleiches bedeuten. Auch die lexikalische Bedeutung in einem Katalog nachzuschlagen hilft nicht weiter, wenn der Zusammenhang nicht gegeben ist.

Diese Erkenntnisse könnten beispielsweise dazu benutzt werden, dem Nutzer einer Suchmaschine Hilfe anzubieten. Bei Eingabe eines Homonyms könnte die Suchmaschine rückfragen, welche Bedeutung der Nutzer meinte und in Abhängigkeit der Antwort nur Texte aus diesem Bereich retournieren.

Auch in der rechnergestützten Übersetzung könnten diese Erkenntnisse Anwendung finden: Wenn in der Zielsprache unterschiedliche Wörter für das (in der Originalsprache) Homonym bestehen, könnte ein Tool den Übersetzungsvorschlag basierend auf den Nachbarn machen, die eine Platzierung des Übersetzungstextes in einer semantischen Karte mit sich bringt (s. Kap. 26, Homonymdisambiguierung: Anwendungsbeispiel).

26 Homonymdisambiguierung: Anwendungsbeispiel

26.1 Überblick

Das grundsätzliche Potential der HM Homonyme zu disambiguieren wird anhand eines Anwendungsbeispiels demonstriert. Die semantische Struktur bereits vorhandener, kategorisierter Texte wird dazu benutzt, die Nachbarschaften neuer Texte zu determinieren und somit deren Kategorisierung – bezogen auf das Homonym – zu prognostizieren.

26.2 Einleitung

Die Disambiguierung von Homonymen ist in der maschinenunterstützten Übersetzung ein häufig anzutreffendes Problem. Wie im vorherigen Kapitel (Kap. 25, Homonyme: Explorationsexperiment Bach/Golf) gezeigt wurde, ist die Hofmethode durch ihre Kontextberücksichtigung grundsätzlich dazu in der Lage, zwischen Homonymbedeutungen zu unterscheiden.

In diesem Kapitel soll ein Anwendungsbeispiel der HM in der Homonym-Disambiguierung demonstriert werden.

Als Übungsbeispiel denken wir uns folgendes Szenario: Von einer deutschsprachigen Textdatenbank sind die englischen Übersetzungen schon erfolgt. Ein Mitarbeiter bekommt einen neuen (deutschen) Text und muss ihn ins Englische übersetzen. Eine Software unterstützt den Mitarbeiter, indem sie eine automatische Übersetzung generiert. Die Software trifft nun anhand einer internen Liste auf ein Homonym («Leiter») und muss entscheiden, ob das Wort eine leitende Person, das Gerät zum Hinaufsteigen oder ein elektrisch leitendes Metall meint, um den entsprechenden englischen Begriff («chief», «ladder» oder «conductor» zu retournieren.

26.3 Vorgehen

Als Datengrundlage diente uns die Patentdatenbank FPO von SumoBrain Solutions[33], die über eine Onlinesuchmaschine komfortabel und schnell durchsuchbar ist. Die Datenbank umfasst US-amerikanische, europäische und japanische Patentbeschreibungen. Neben diversen Angaben zur Person/Firma des Patentinhabers enthält jeder Patenteintrag ein Abstract und die genaue Beschreibung des eigentlichen Patentes. Wir beschränkten uns auf die Abstracts, da diese den vollumfänglichen, semantischen Raum eines jeden Eintrages umfassen sollten.

Schritt 1

Wir gaben als Suchbegriff «Leiter» ein, worauf 62'273 Treffer retourniert wurden. Von den ersten 60 übernahmen wir die Abstracts in unsere eigene Datenbank. Um eine englische Übersetzung des Wortes «Leiter» zu simulieren, wiesen wir manuell jedem Text – je nach Verwendung von «Leiter» – die Kategorie «ladder» oder «conductor» zu (Leiter im Sinne von Führungsperson kam nicht vor). Als Keywords nutzten wird das Tagcloud-API (s. Kap. 16, Tagcloud-Verfahren von Semager). Die TargetWords wurden bestimmt, behoft und die semantische Karte erstellt. Diese Karte soll in unserem einfachen Szenario die Ausgangskarte sein – sie beinhaltet quasi das Referenzmaterial, mit welchem die neu hinzukommenden Texte verglichen werden. In einem realen Szenario würde man diese Ausgangskarte selbstverständlich mit ausgewählten Texten berechnen, die in ihrer Semantik besonders repräsentativ für die Homonym-bedeutungen wären.

Schritt 2

Dann wurde ein neuer Text aus der Patentdatenbank gezogen, der das Wort «Leiter» enthielt. Dieser Text repräsentiert nun das neu zu übersetzende Patent. Der Hof des Wortes «Leiter» wird bestimmt und mit den übrigen Texten verglichen (durch die Normierung nach TotalTargetWords (s. Kap. 5, Normierung der Textähnlichkeitswerte: SharedTargetWords vs. TotalTargetWords) fallen die unterschiedlichen Hofhäufigkeiten nicht ins Gewicht) und eine neue Karte erstellt. Schliesslich wurden vom neuen Text die drei nächsten Nachbarn und deren Kategorien bestimmt. Diejenige Kategorie, welche mindestens zwei Mal vorkommt, ist der Prädiktor für die Kategorie des neuen Textes.

Schritt 2 wurde fünf Mal mit weiteren Texten wiederholt. Wir nahmen jeweils die nächstfolgenden Texte aus der Resultatsliste aus Schritt 1, achteten aber darauf, dass «Leiter» drei Mal im Sinne von Gerät und drei Mal im Sinne von Elektroleiter vorkamen.

33 http://www.freepatentsonline.com

26.4 Resultate

In Abbildung 127 ist die Ausgangskarte zu sehen. Gelb eingefärbt sind die Patente, in denen das Wort «Leiter» im Sinne von «elektrischer Leiter» (conductor) vorkommt, rot eingefärbt die Patente, welche das Arbeitsinstrument «Leiter» (ladder) zum Thema haben. Die beiden Wortbedeutungen separieren mehr schlecht als recht, trotzdem sind Bedeutungsbereiche vorhanden. Die conductor-Texte bilden einen Bereich in der Kartenmitte und einen peripheren Bereich rechts. Der Bereich der ladder-Texte wird durch conductor-Texte unterteilt, der sich wie einen Keil dazwischen schiebt.

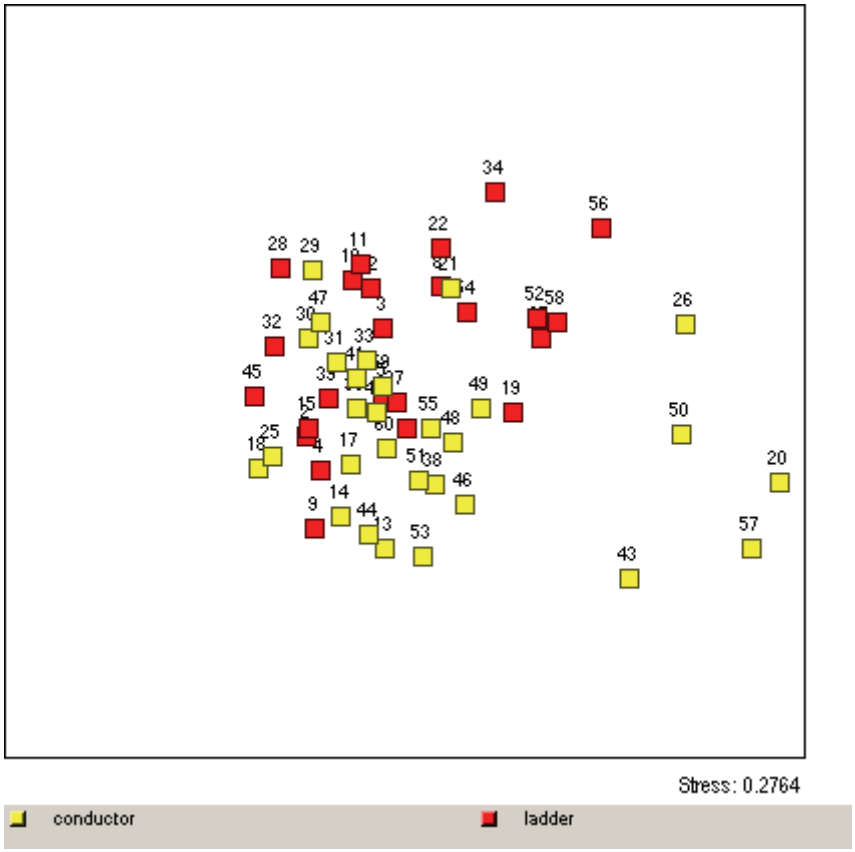


Abbildung 127: Die sog. Ausgangskarte zeigt die 60 retournierten Patente. Eingefärbt sind die beiden Bedeutungskategorien elektrischer Leiter (gelb) und Gerät (rot).

In Abbildung 128 sind exemplarisch zwei Karten dargestellt, in welche je ein neues Textitem (weiss eingefärbt) hinein gerechnet wurde. Bei der linken Karte determinieren die Nachbarn die Kategorie des neuen Items korrekt (conductor). Bei der rechten Karte gibt der nächstliegende Nachbar die Kategorie des neuen Textes zwar korrekt an (ladder), jedoch nicht die beiden folgenden Nachbarn, somit erfolgte die Klassifizierung mit dem verwendeten Verfahren falsch.

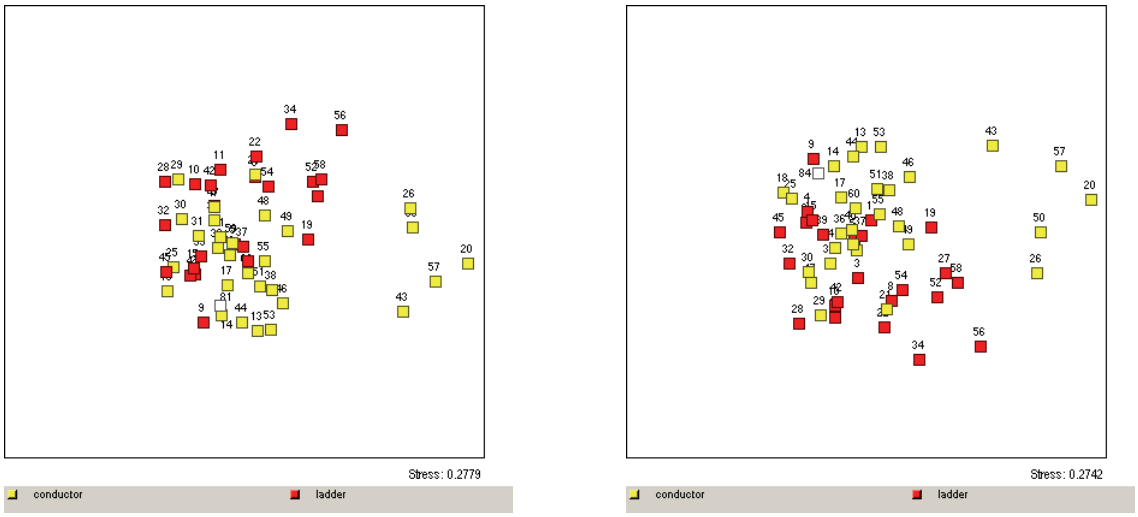


Abbildung 128: Neu berechnete Karten mit zusätzlichen unkategorisierten Texten (Id 81 links, bzw. Id 85 rechts; beide weiss eingefärbt).

Von den sechs Texten, deren Kategorie errechnet werden musste, wurden fünf korrekt berechnet (s. Tab. 13).

Tabelle 13: Von sechs neuen Texten konnte die Kategorie in fünf Fällen korrekt berechnet werden, ein Mal nicht (fett gedruckt).

Id	Nachbarn (c=conductor, l=ladder)	errechnete Kategorie	richtige Kategorie
80	conductor/ladder/ladder	ladder	ladder
81	conductor/conductor/ladder	conductor	conductor
82	conductor/ladder/conductor	conductor	conductor
83	conductor/conductor/ladder	conductor	conductor
84	ladder/conductor/conductor	conductor	**ladder**
85	ladder/ladder/conductor	ladder	ladder

26.5 Diskussion

Dieses sehr einfach gestrickte Experiment demonstriert, wie die Hofmethode im Bereich Homonym-disambiguierung prinzipiell von Nutzen sein kann. Von sechs Texten konnten in fünf Fällen die Zielübersetzung des Homonyms «Leiter» korrekt prognostiziert werden.

Dieses Experiment hat nur demonstrativen Charakter. Es wäre falsch, daraus den Schluss zu ziehen, dass die HM eine Homonymerkennungsquote von 83.3% produziert. So basiert die Ausgangskarte auf – was die Homonymverwendung anbelangt – zufälligen Texten. Geschickter wäre es, die Texte so auszuwählen, dass die semantische Verwendung von «Leiter» möglichst prägnant zum Ausdruck kommt. Die diffuse Kategorisierung der Karte könnte zudem verbessert werden, indem mehr Items aufgenommen würden. Auch wurden nur die Abstracts behoft, nicht aber die detaillierten Patentbeschreibungstexte.

Auch das Verfahren zur Kategoriedetermination wurde nicht weiter elaboriert: Weder spielen die Distanzen zum Zielobjekt eine Rolle, noch die Nachbarschaftsstruktur (Clusterung oder periphere Lage). Die errechnete Zielkategorie könnte zudem mit einem Wahrscheinlichkeitswert versehen werden, um ihre Aussagekraft besser einschätzen zu können.

Weiter wäre es prüfenswert zu untersuchen, ob es genügen würde, nur die Homonyme zu behofen, statt der gesamten Texte mittels einer Keywordliste.

Zur Effizienzsteigerung könnte der neue Zieltext nur mit den Paarähnlichkeitswerten der Ausgangskarte verglichen werden; somit würde die rechenintensive NMDS wegfallen. Ob darunter die Qualität der Homonymerkennung signifikant leiden würde, bleibt zu untersuchen.

Die Anwendungsmöglichkeiten einer Homonymdisambiguierung sind vielfältig. Überall, wo Text nach semantischen Kriterien verarbeitet wird, sind Homonyme potenzielle Fehlerquellen. Gerade in der maschinenunterstützten Übersetzung ist aber ein grosses Bedürfnis nach einem derartigen Werkzeug auszumachen.

27 Experiment DSM-IV und Diagnostik

27.1 Überblick

Es wird untersucht, ob die Hofmethode Ähnlichkeiten in Beschreibungen psychischer Störungsbilder sinnvoll berechnen kann. Weiter wird untersucht, ob diese Karten dazu geeignet sind, Fallbeispiele mit diagnostischen Hinweisen zu versehen. In diesem explorativen Experiment spielt die Textherkunft leider die dominierende Rolle und verhindert somit semantisch brauchbare Karten.

27.2 Einleitung

In diesem explorativen Experimenten tasten wir uns an die Frage heran, ob sich die HM dazu eignet, Hilfe bei der Diagnose von psychischen Störungen zu bieten. Die Thematik ist äusserst komplex und würde genügend Untersuchungsmaterial für eine eigene Dissertation bieten. Wir werden in diesem Kapitel sehr skizzenhaft vorgehen und nur grob illustrieren, wie ein möglicher Einsatz der HM in der Diagnostik aussehen könnte und wo Probleme erwartet werden müssen.

Als vorbereitenden Schritt stellt sich die Frage, inwieweit die HM Ähnlichkeiten von Beschreibungen psychischer Störungen erfassen kann. Dazu behofen wir eine Auswahl psychischer Störungen und vergleichen die resultierende NMDS-Struktur mit einem Expertenmodell. Als primäre Textgrundlage dient das DSM-IV (Sass, 1998), welches die Störungen in ausführlicher Weise beschreibt (im Gegensatz zum ICD-10 (Dilling, 1994), in dem die Störungen nur stichwortartig beschrieben werden). Die kategoriale Beschreibung von psychischen Krankheiten ist in der Psychologie ein umstrittenes Thema (für eine Übersicht s. Egli, 2006). Für gewisse Krankheitstypen eignete sich ein dimensionales Modell besser. Im erwähnten Artikel kombinieren Egli et al. ein kategoriales mit einem dimensionalen Modell, indem sie 20 Psychotherapeuten Ähnlichkeiten von 21 psychischen Störungen einschätzen lassen und mit diesen Ähnlichkeitsurteilen eine NMDS berechnen. Die resultierende Karte weist einerseits Bereiche auf, die mit der Klassifikation des ICD-10 übereinstimmen, andererseits gibt es Überlappungen zwischen Kategorien, die inhaltlich durchaus der Realität entsprechen. Diese Karte dient uns als Expertenmodell.

In einem weiteren Schritt untersuchen wir, wie sich Beschreibungen von Patienten (Fallbeispiele) in diese Struktur einfügen. Sollten die Diagnose dieser Patienten den Störungsbildern in der unmittelbaren

Nachbarschaft entsprechen, könnte ein auf diesen Erkenntnissen programmiertes Tool potenziell Hilfe bei der Diagnose von Patientenbeschreibungen anbieten.

27.3 Vorgehen

Wir übernahmen die Auswahl der psychischen Störungen aus dem Expertenmodell. Da diese nach der ICD-10-Klassifikation codiert waren, suchten wir die jeweils besten Entsprechungen im DSM-IV-Modell. Die Texte wurden gescannt und einer automatischen Texterkennung (OCR) unterzogen. Metainformationen wurden nicht aufgenommen (Kapitelüberschriften oder Störungscodes). Das DSM-IV ist zum Teil aufbauend, d. h. zu verschiedenen, ähnlichen Störungsbildern wird zuerst eine allgemeine Erläuterung gegeben und erst im spezifischen Störungsbeschrieb wird auf die Besonderheiten eingegangen. In diesen Fällen haben wir den allgemeinen Teil für jedes spezifische Störungsbild in den Beschrieb mit einbezogen. Dieser Umstand darf nicht vergessen werden, wenn bestimmte Textpaare eine sehr hohe Ähnlichkeit aufweisen[34].

Die Fallgeschichten wurden aus drei Büchern zusammengetragen (Lieb, 2009; Freiberger, 1999; Dilling 2000). Bei den Fallgeschichten wurde nur die Anamnese erfasst, also keine Vorgeschichten oder Laborbefunde.

Die Keywords wurden mit der Wortfrequenzmethode (s. Kap. 14, Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient) bestimmt (Abbruchgrenze für TargetWords: 0.02). 80 Keywords resultierten, mit denen die Texte behoft wurden (unter Einhaltung eines TargetWord-Koeffizienten zwischen 0.02 und 0.14).

Mit den DSM-IV-Texten wurde eine Karte gerechnet und mit dem Expertenmodell verglichen. Da es nicht dieselben Items waren und eine unterschiedliche Anzahl, konnte keine Prokrustes-Transformation gerechnet werden. Der Vergleich beschränkt sich auf die visuelle Beurteilung der groben Struktur.

Schliesslich wurde eine gemeinsame Karte erstellt, in der alle Texte – die DSM-IV-Texte und die Fallbeispiele – vertreten waren.

34 Dieser Umstand würde eigentlich dem ÜK entgegenkommen: Gleiche Textpassagen erhöhen den gemeinsamen Ähnlichkeitswert stark. Wird über alle DSM-IV-Texte der ÜK gerechnet, resultiert aber eine Karte, die qualitativ deutlich schlechter ist, als die HM-Karte (s. Anhang 6).

27.4 Resultate

Der Vergleich mit der Expertenkarte (Abb. 129) zeigt ein vielversprechendes Bild: Die ICD-10-Kategorien sind auch in der DSM-IV-Karte, die durch die HM errechnet wurde, als Bereiche vorhanden: Sehr deutlich bei den Störungen der ICD-10-Kategorien F2 (orange), F3 (olivgrün), F6 (blau) und F10

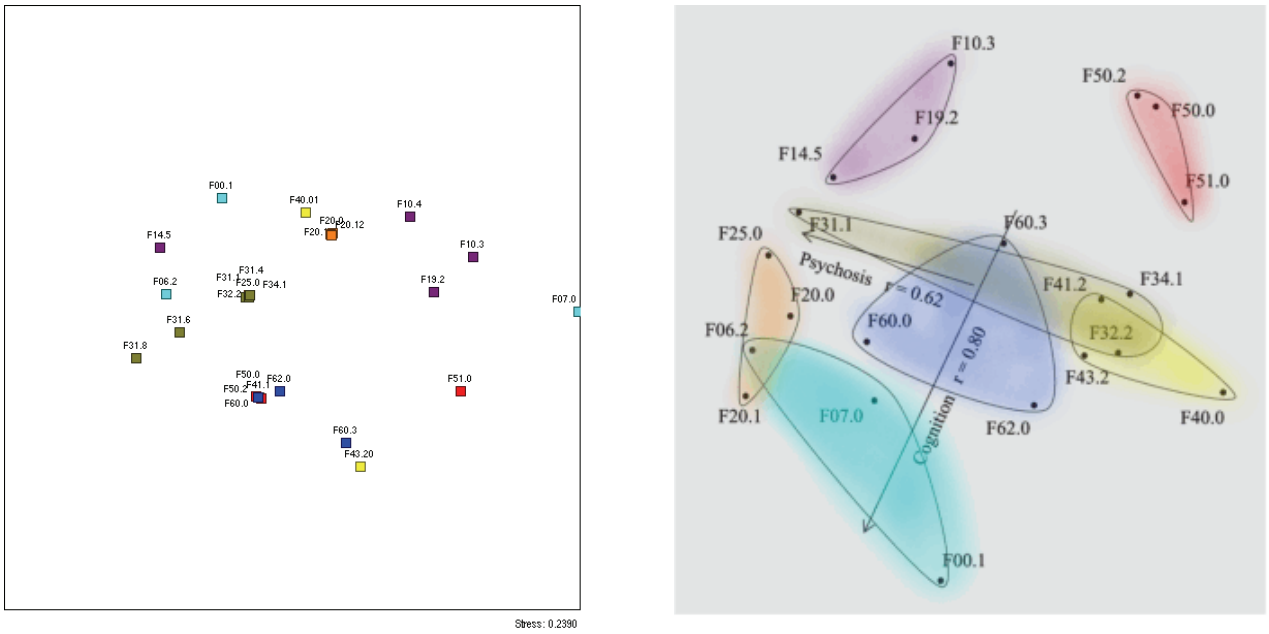


Abbildung 129: Links:Texte des DSM-IV nach der Behofung; rechts: Expertenkarte. Die Einfärbung bezieht sich auf die Einteilung der Störungsbilder nach ICD-10.

(violett), weniger deutlich bei F0 (türkis) und F5 (rot), unklar bei F4 (gelb) . Die Bedeutung der Farbkodierung ist in Abbildung 130 ersichtlich. Die Anordnung der Bereiche weist Gemeinsamkeiten mit derjenigen der Expertenkarte auf, jedoch sollte hier nicht zu viel hinein interpretiert werden. Innerhalb der Bereiche gibt es einige Abweichungen, beziehungsweise Texte, die nicht innerhalb ihrer Kategorie liegen. Allerdings darf nicht ausser Acht gelassen werden, dass wir hier Texte des DSM-IV vor uns haben, die nach dessen Logik geschrieben wurden. Die ICD-10-Einteilung ist in weiten Bereichen identisch, allerdings nicht in allen. Färben wir dieselbe Karte nach den Kategorien des DSM-IV ein, ergibt sich ein etwas konsistenteres Bild (Abbildung 130).

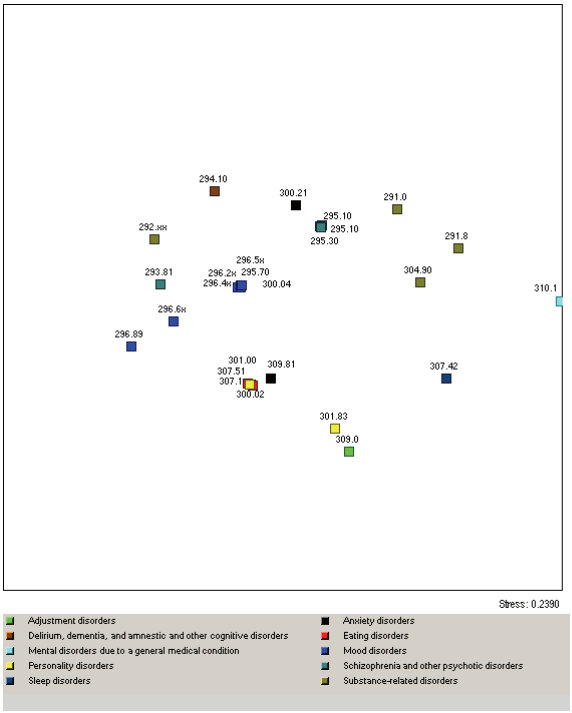


Abbildung 130: Die DSM-IV-Karte nach der Einfärbung nach den DSM-IV-Kategorien.

Wir werden hier nicht tiefer in die Analyse der Karten einsteigen, sondern begnügen uns mit der Erkenntnis, dass sinnvolle Karten der Symptombeschreibungen aus dem DSM-IV mithilfe der HM grundsätzlich möglich sind.

Die Integration der Fallgeschichten in die DSM-IV-Karte zeigt Abbildung 131. Es ist rasch erkennbar, dass die Karte keine einheitlichen Bereiche bildet. Die Texte sind auf den ersten Blick zufallsverteilt.

Färben wir dieselbe Karte nach der Herkunft der Texte ein, zeigt sich doch eine Ordnung, leider keine erwünschte (Abb. 132). Die Integration der Fallbeispiele klappt eindeutig nicht. Die DSM-IV-Texte bleiben unter sich und auch die Fallbeispiele mischen sich nicht vollständig. Hier zeigt sich eine Limitation der HM: Wenn die Sprachstile zu unterschiedlich sind und dazu ein Stil in einem markanten Teil der Texte verwendet wird, kann es passieren, dass dieser Sprachstil zum dominierenden Faktor in der Ähnlichkeits-berechnung wird.

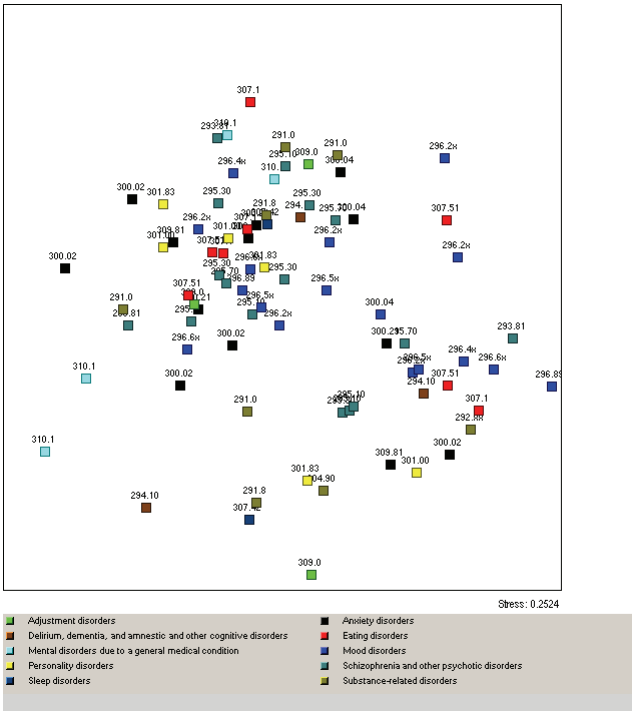


Abbildung 131: Die Integration der Fallbeispiele in die DSM-IV-Karte produziert nicht die semantisch gewünschte Karte.

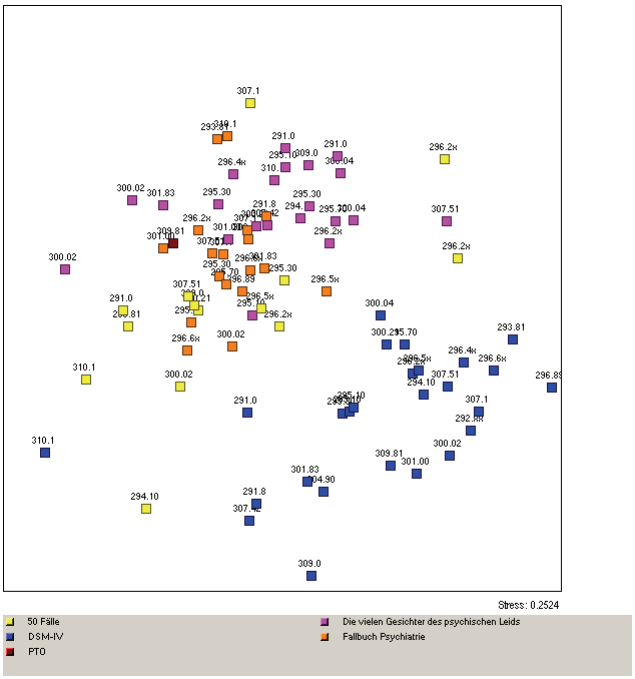


Abbildung 132: Die Einfärbung der gemeinsamen Karte nach Textherkunft zeigt den dominierenden Faktor der Ähnlichkeitsberechnung.

Auch wenn man die DSM-IV-Texte weg lässt und nur die Fallgeschichten behoft, bleibt die Textherkunft der dominierende Faktor in der Ähnlichkeitsberechnung (s. Abb. 134).

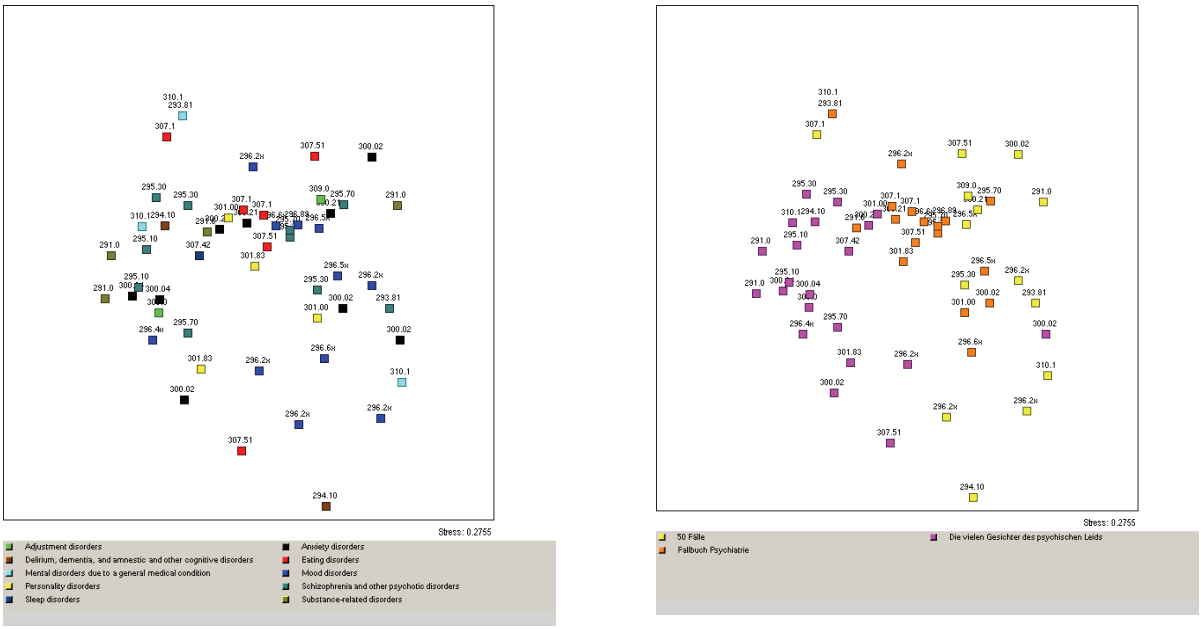


Abbildung 133:Auch ohne die dominierenden DSM-IV-Texte bilden die Fallgeschichten keine semantisch brauchbare Karte. Links sind die Fallgeschichten nach DSM-IV-Kategorie eingefärbt, rechts nach der Textherkunft. Diese bleibt der dominierende Faktor.

27.5 Diskussion

Innerhalb der DSM-IV-Störungsbeschreibungen erzeugt die HM semantisch durchaus sinnvolle Karten. Werden jedoch Fallbeispiele aus anderen Quellen hinzugenommen, gerät der Schreibstil zum dominierenden Ordnungsfaktor. Die ursprüngliche Idee, Fallbeispiele in die DSM-IV-Karte reinzurechnen, um Hinweise auf die Diagnose zu bekommen, funktioniert also nicht.

Für einen praktischen Einsatz müsste weitere Forschungsarbeit geleistet werden. Der Einfluss des Schreibstils wäre zu eliminieren, sei es rechnerisch oder durch eine strukturiertere Diagnose.

Es ist anzumerken, dass die Auswahl der Keywords ja mit der Wortfrequenzmethode erfolgte. Diese erzeugt eine grobe Kategorisierung mithilfe des ÜK. Der ÜK ist jedoch besonders anfällig für unterschiedliche Schreibstile.

28 Semantische Strukturierung eines Diskussionsraums

28.1 Überblick

Die HM wird eingesetzt, um die Texte einer Datenbasis zu strukturieren. Ziel ist es, dem Leser eine rasche Orientierung zu ermöglichen.

28.2 Einleitung

Kann die HM ein Konvolut von Diskussionstexten semantisch so strukturieren, dass ein menschlicher Leser einen Überblick der vorkommenden Themen erhält? Dass er erkennen kann, aus welchen Facetten ein Thema besteht oder wie kontrovers es behandelt wird? Und wie die Themen strukturell zueinander stehen?

Die Anwendungsmöglichkeiten sind vielfältig – eine davon ist das Online-Diskussionsforum. Anlässlich der Stadtdebatte (s. Anhang, Kap. 2.1, Beschreibung des Webforums Stadtdebatte) hatten wir Gelegenheit, diese Fragen zu prüfen.

28.3 Vorgehen

Drei Vorgehensweisen werden erprobt:

Strukturierung innerhalb eines Forums

Können aus einer Auswahl von Texten innerhalb eines Forums die diskutierten Themen wiedergegeben werden?

Strukturierung mehrerer Foren

Wie verhalten sich die Texte verschiedener Foren innerhalb einer Karte?

Strukturierung hinsichtlich eines Ausgangstextes

Wie strukturiert sich der Raum, wenn keine zufälligen Items ausgewählt werden, sondern die ähnlichsten Texte zu einem bestimmten Ausgangstext?

28.3.1 Beschränkung der Daten

In einer Karte lassen sich etwa bis 60 Items unterbringen, ohne dass die Lesbarkeit der Itemtitel allzu stark leidet. Um einen Textkorpus von 2'000 Items zu strukturieren, muss also eine Auswahl getroffen werden.

Eine Zufallsauswahl von 60 Texten aus dem gesamten Datenbestand dürfte sehr heterogen ausfallen, sogar, wenn die Auswahl auf ein einziges Forum beschränkt wird: Viele der 2'000 Beiträge sind Antworten auf einen Startbeitrag – entweder unterstützend, im Sinne von «ja, das geht mir auch so», oder aber entgegengesetzt: «finde ich gar nicht»; jedenfalls handeln sie vom gleichen Thema, wie der Startbeitrag. Insgesamt waren 337 Texte Startbeiträge. Da sich Themen teilweise wiederholten, können wir mit höchstens 300 unterschiedlichen Themenbereiche rechnen. Bei den fünf Foren macht das 60 Themenschwerpunkte pro Forum (wobei zu beachten ist, dass die Beiträge unterschiedlich auf die Foren verteilt waren). Nimmt man eine Zufallsauswahl von 60 Texten aus einem Forum, ist also keine Clusterbildung zu erwarten, jedoch sollten die Beiträge immerhin thematisch zueinander in einer sinnvollen Relation stehen.

Die Auswahl der Keywords erfolgte bei allen Berechnungen in diesem Kapitel durch die Wortfrequenzmethode (s. Kap. 14, Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient), wobei nur die ersten 50 Keywords genommen wurden. Das Hofgewicht (s. Kap. 6, Hofgewichtung) wurde auf 10 festgelegt.

28.4 Resultate

28.4.1 Strukturierung innerhalb eines Forums

Aus dem Forum «Wie soll sich Zürich baulich verändern?» wurden per Zufall 60 Texte ausgewählt und behoft. Sieben Texte wiesen keine Keywords auf und wurden aus der Auswahl wieder entfernt, die restlichen 53 Items wurden skaliert. Damit die resultierende Karte besser interpretierbar ist, wurden die Texte ganz grob mittels einem, manchmal auch mehreren, Stichwörtern charakterisiert und damit angeschrieben.

In Abbildung 134 ist die resultierende Karte zu sehen. Rechts unten ist ein thematischer Schwerpunkt mit Fokus auf «verdichten» und «Architektur». Im Raum frei verteilt sind diejenigen Beiträge mit dem

Thema «Grünflächen». Eine grobe Struktur ist zwar erkennbar, jedoch könnte man anhand der Karte nicht zielgerichtet einen bestimmten Text finden. Ein möglicher Nutzen wäre höchstens auf einer sehr hohen Abstraktionsebene zu finden, indem die Bereiche mittels einer Clusteranalyse identifiziert würden und mit den häufigsten Wörtern gelabelt. Diese Idee wird hier jedoch nicht weiterverfolgt.

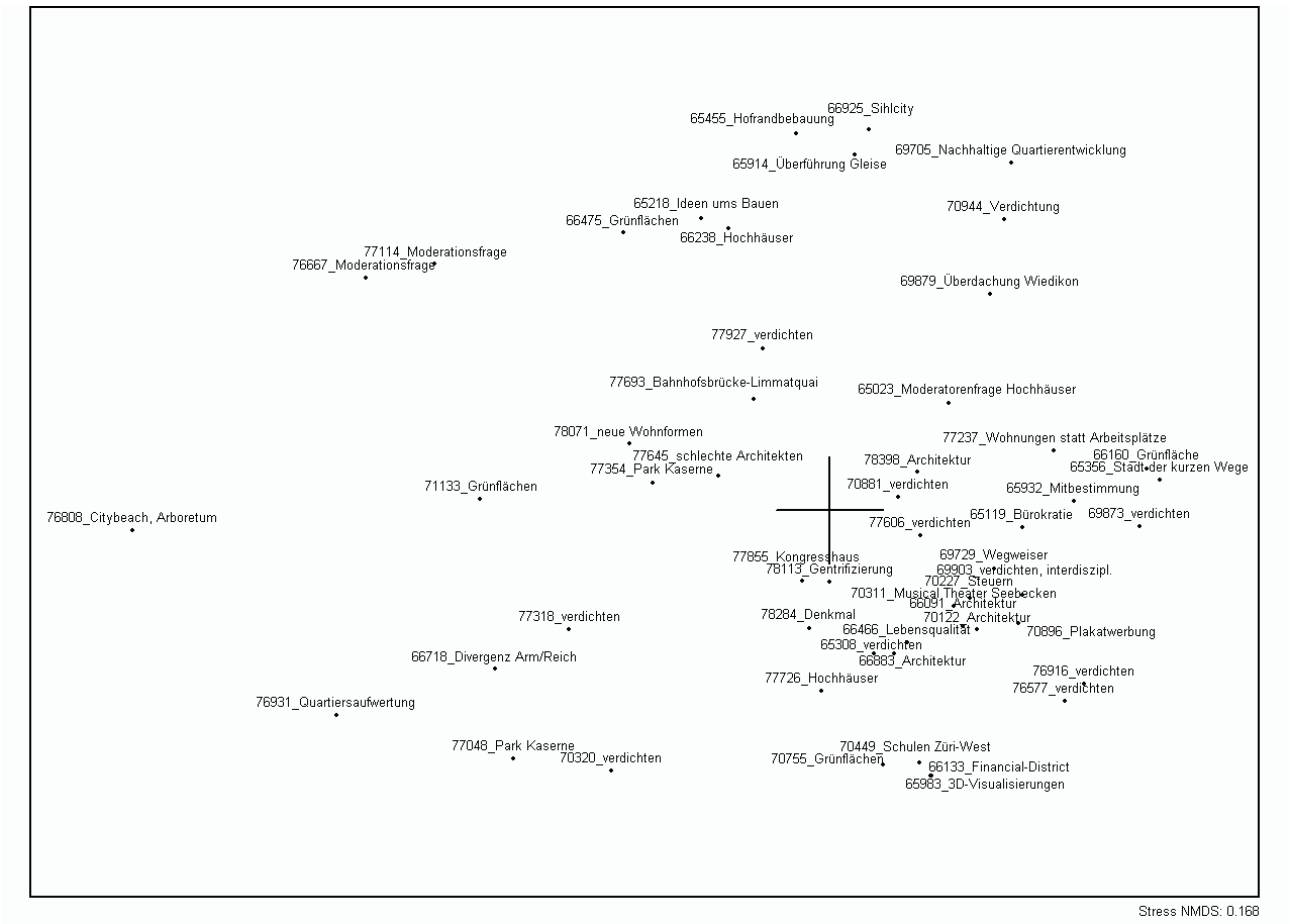


Abbildung 134: 60 Texte aus dem Forum «Bauen». Eine grobe Struktur ist ersichtlich.

28.4.2 Strukturierung mehrerer Foren

Was passiert, wenn die Beiträge mehrerer Foren gemischt werden? In einem Versuch wurden aus den drei Foren «Mobilität», «Bauen» und «Grenzen» je 20 Texte per Zufall bestimmt, behoft und skaliert (s. Abbildung 135). Die drei Foren separieren eindeutig nicht, aber ist das inhaltlich sinnvoll?

Zunächst die grobe Struktur der Karte: Rechts ist ein inhaltlicher Schwerpunkt zum Thema Bauen, Verdichtung, Architektur. Darunter sind Texte zum Thema Wohnen zu finden. Links unten ist ein Schwerpunkt mit den Themen Velofahrer, Tempo 30 und autofrei, ebenso oben-mittig und oben-links. Mitte-links ist recht divers.

Betrachten wir nun das Forum «Grenzen». Item 66544 (in der Mitte links) handelt von einer Raumplanungsstrategie und hat wenig mit den anderen Items zu tun, ebenso wenig wie Item 70512, schräg rechts darüber, das einen vorherigen Schreiber der Naivität bezichtigt. Aber die Items 65509, 70023 und 76655 handeln alle drei von Raumplanung; sie sind im Themenschwerpunkt «Bauen/Verdichten» nicht schlecht aufgehoben, dürften aber näher beieinander sein. Item 76796 mit Thema «Wohnen» ist richtig platziert.

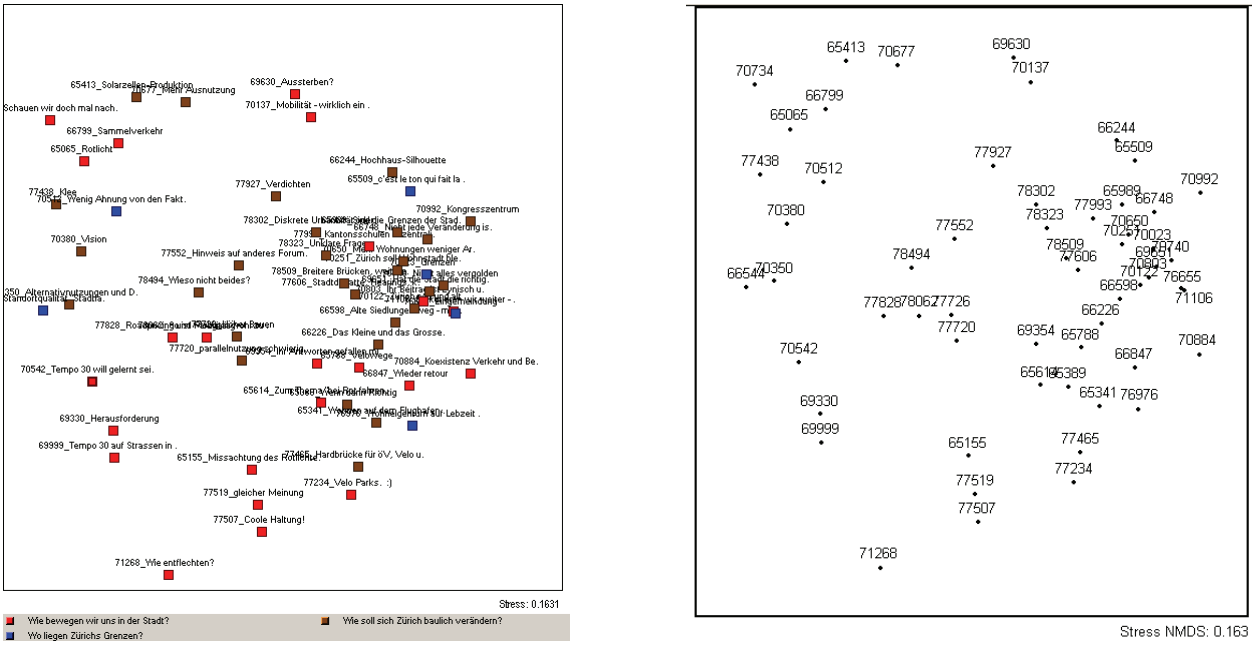


Abbildung 135: Drei Foren in einer Karte, wieder ist nur grobe Struktur erkennbar. Die Texte sind links mit den Betreffs der Verfasser angeschrieben, rechts ist dieselbe Karte, jedoch sind nur die Text-Ids angegeben.

Das Forum «Mobilität» ist zu stark verteilt. Es ist inhaltlich nicht nachvollziehbar, warum die oberen Text nicht bei den anderen in der unteren Kartenhälfte sind. Das Forum «Bauen» konstituiert vor allem den Themenbereich «Verdichten», auch das Thema «Wohnen» etwas unterhalb der Mitte. Die Makrostruktur ist durchaus sinnvoll, jedoch ist die Intraclusterstruktur nicht verlässlich.

28.4.3 Strukturierung hinsichtlich eines Ausgangstextes

Als Ausgangstext diente eine markante Aussage aus dem Forum «Mobilität», in der es um gefährliche Velowege geht. Dann wurde für alle anderen Texte der Überlappungskoeffizient mit diesem Ausgangstext berechnet und diejenigen 60 Texte mit den höchsten Paarwerten behoft und skaliert. Das Ergebnis ist in Abbildung 136 zu sehen. Der Ausgangstext ist weiss eingefärbt.

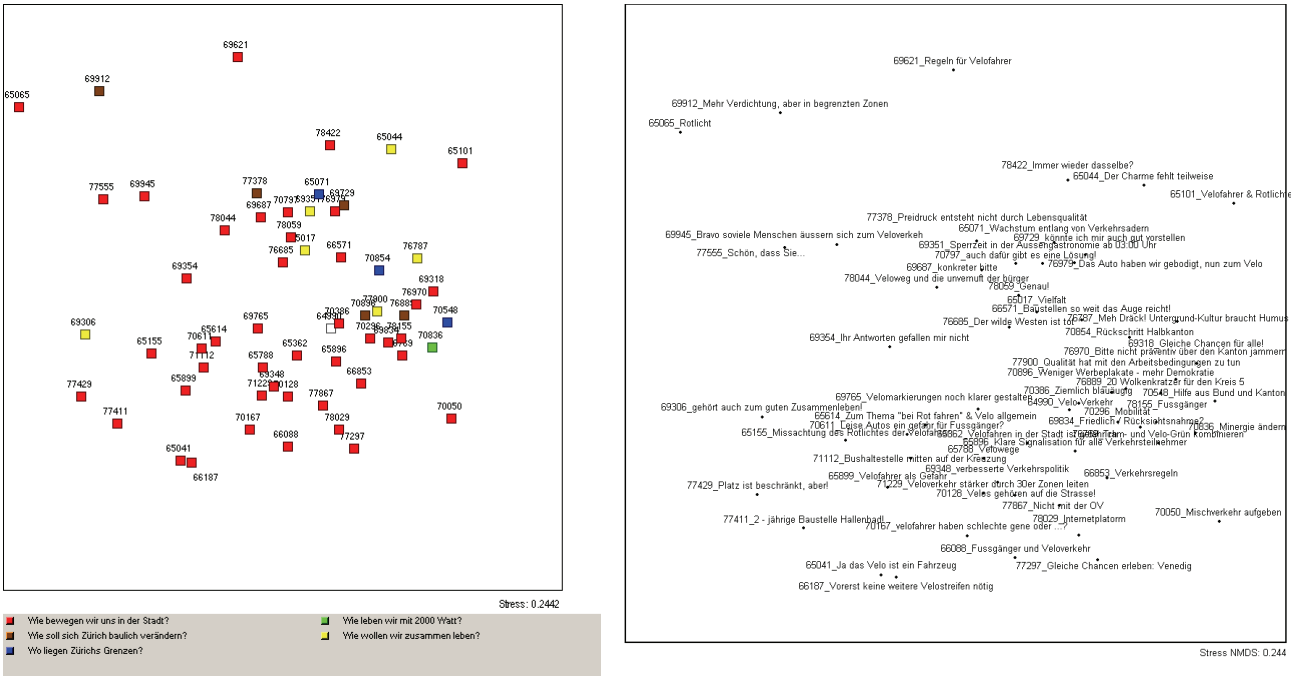


Abbildung 136: Diejenigen 60 Items mit dem höchsten ÜK zu Item 64990 (weiss) wurden skaliert. Links sind die Foren eingefärbt. Rechts sind die Beitragstitel angegeben.

In unmittelbarer Nachbarschaft des Ausgangstextes sind Items mit äusserst ähnlichem Tonfall und Inhalt. Folgende Auszüge verdeutlichen dies:

Ausgangstext (Item 64990):

> Warum hören die markierten Velowege an den gefährlichsten Stellen auf? In der ganzen Stadt! Warum bringt es die Stadt Zürich nicht fertig, die Velofahrer bei Lichtsignalen "rechts-abbiegen" zu lassen. Ist in anderen europäischen Städten seit jahrzehnten üblich.

Item 65362:

> Ich finde jedoch, dass Velofahren in der Stadt schon gefährlich ist. Seit ich mit Kindern unterwegs bin, plane ich meine Routen durch die Stadt vorsichtiger und fahre nicht mehr einfach von A nach B. Es ärgert mich dass ich mein Engagement für die Umwelt mit einem höheren Verletzungs- oder Todesrisiko (insbesondere auch meiner Kinder!) bezahlen muss.

Item 65896:

> Ich wohne ebenfalls in der Stadt Zürich, bin innerhalb der Stadt fast ausschliesslich mit dem Velo unterwegs und stelle oftmals fest, dass an vielen Stellen wo die Verkehrswege von Auto/Velo/Fussgänger aufeinandertreffen, diese nicht eindeutig gekennzeichnet sind. So ist für Fussgänger/Velofahrer ausser sporadischen Markierungen am Boden nicht immer klar gekennzeichnet wo der Velo/Fussgängerbereich beginnt und wo dieser endet.

Item 66853:

> Unser Verkehr in der Stadt Zürich ist vielfältig und komplex geworden;nun haben wir stets Stau und Chaos! Autos, Off Roads, Lastwagen,Bus und Trams, Roller, Velos, Kinderwagen, Trottinetts,Fussgänger;die Schnellen, die Langsamen, die die Musik hören und nicht den Verkehr, die die stets telefonieren und die kleineren gefährden, die dem Bus kein Vortritt lassen,….

Item 69765:

> ich bin erst seit kurzem mit dem E-Bike unterwegs in der Stadt mir fällt auf, dass wenn man nicht kundig ist, häufig die Abzweigungen der Velowege verpasst; entweder muss man gleich eine Vollbremsung machen oder man schiesst am Ziel vorbei. Bei längeren Velowegen wie z.B. vom Paradeplatz über den Schanzengraben ist auf der Brücke dies viel schlechter markiert als am See unten, dort ist es gut ersichtlich, dass Fussgänger und Fahrrad sich den Platz teilen müssen und es funktioniert dank der klaren Signalisierung nicht schlecht.

Item 65788:

> ganz meine Meinung, aber da scheint der Wille zu fehlen dies auch so umzusetzen, auch in Winterthur ist es angeblich möglich klar signalisierte Velowege zu markieren, und auch an gefährlichen Stellen in roter Farbe!

Item 70128:

> Dabei fällt jedem Velofahrer in Zürich auf, dass man durch die Markierung der Velowege dazu erzogen wird, aufs Trottoir auszuweichen.

Item 70296:

> Die Stadt Zürich soll noch viel Fussgänger- und Velotauglicher werden.

Item 70386 schlägt dann die Brücke zu wirtschaftlichen Aspekten (Item 70896: Plakatwerbung, 77900 Arbeitsplätze).

Item 70386:

> Wenn ich dieses Hick Hack um die armen Radfahrer lese - wohlgemerkt, ich bin selber überzeugter Radler und ich habe mein ganzes Leben in der Stadt Zürich gewohnt - bekomme ich das Gefühl, die Stadt floriere ausschliesslich dank den Radfahrern.

Im unteren Teil der Karte sind weitere Beiträge zu Velowegen und -fahrern, im oberen Teil eher zu Rotlichtern und Regelungehorsam.

28.5 Diskussion

Die beiden ersten Beispiele zeigen, dass mit zufällig ausgewählten Texten eines grossen Diskussionsforums zwar eine grobe Struktur gebildet werden kann, jedoch nichts darüber hinaus. Der praktische Nutzen bleibt bei dieser knappen Strukturierung beschränkt. Die Karten könnten allenfalls genutzt werden,

um über eine Clusteranalyse Themenschwerpunkte zu finden, die unter Auswertung der häufigsten Wörter gelabelt werden könnten.

Das dritte Beispiel, bei dem die ÜK-ähnlichsten Texte zu einem Ausgangstext skaliert werden, liefert spannendere Ergebnisse: Basierend auf der Ähnlichkeit zu einem Ausgangstext wird eine Anzahl von Texten aus der Datenbank gezogen und skaliert. Die resultierende Karte präsentiert diese Texte so nach ihrer gegenseitigen semantischen Ähnlichkeit, dass Texte, die in ihrem Duktus sehr ähnlich dem Ausgangstext sind, leicht identifiziert werden können.

29 Kategorisierung: Eingliederung neuer Elemente in eine bestehende Taxonomie

29.1 Überblick

Die HM wird in diesem Kapitel benutzt, um die von einem menschlichen User zugewiesene Kategorie eines neuen Elementes zu verifizieren, indem das Element semantisch mit bestehenden Texten und deren Kategorien verglichen wird.

29.2 Einleitung

Gerade im Web-Umfeld ist es oft nötig, neue Elemente in eine bestehende Taxonomie aufzunehmen. Man denke an Diskussionsforen, schwarze Bretter, Versteigerungsplattformen oder überhaupt Online-Shops. Die Kategorie-Zuweisung kann durch den Dienstanbieter erfolgen (bspw. Amazon) oder durch den User (Ebay/Ricardo, Diskussionsforum). In allen Fällen kann eine Vielzahl von Fehlern passieren: Falsche Kategorisierungen, weil das neue Element ambivalente Facetten aufweist; unterschiedliche Sichtweisen der Kategorisierer; Unkenntnis der vorhandenen Kategorien oder des neuen Elementes; etc.

Da die Hofmethode für sich in Anspruch nimmt, semantische Ähnlichkeiten aufzuspüren, wird in diesem Kapitel der Anwendungsfall eines Webforums untersucht: Die HM vergleicht die von einem Erfasser zugewiesene Kategorisierung seines Textbeitrages mit den bereits vorhandenen Einträgen. Falls der Text des Beitrags eher auf eine andere Kategorie passte, würde dies dem User zurückgemeldet.

Als Datengrundlage dienen die Texte der Stadtdebatte (s. Anhang, Kap. 2.1, Beschreibung des Webforums Stadtdebatte). Jeder Autor musste seinen Beitrag in einem der fünf vorgegebenen Foren posten. Der Beitrag konnte einen neuen Thread starten (Startbeitrag) oder eine Antwort/ein Kommentar zu einem bestehenden Beitrag sein. Um den SemanticMapper exemplarisch zu testen, werden die 10 zuletzt geposteten Startbeiträge herangezogen.

29.3 Vorgehen

Das Verfahren ist vom Prinzip her sehr einfach: Der zu untersuchende Text wird mit dem gesamten, bestehenden Datenbestand verglichen und die Foren der 20 ähnlichsten Texte werden ausgezählt. Die Ähnlichkeit kann dabei aus dem ÜK oder der semantischen Ähnlichkeit nach HM bestehen.

Variante A (nur ÜK)

Aus einer vorberechneten Dreiecksmatrix mit den ÜK aller Texte werden die 20 Items mit den höchsten Paarwerten herausgesucht und deren Foren ausgezählt. Das Finden von beliebig vielen Paarwerten in einer DEM ist übrigens nicht rechenintensiv und dauert keine Zeit.

Variante B (HM + ÜK)

Die 20 ÜK-ähnlichsten Items werden behoft und skaliert.

Variante C (nur HM)

Der zu untersuchende Text wird mit allen Texten der Datenbank mittels Hofmethode verglichen.

Ein semantischer Vergleich mit der Hofmethode ist bei 2'000 Texten rechenintensiv. Effizienter ist es, aus dem Textkorpus die 100 ähnlichsten Items mittels ÜK heraus zu suchen, dann mit diesen 100 Texten einen semantischen Hofvergleich mit dem Startitem zu machen und schliesslich die 20 Texte mit der grössten (semantischen) Ähnlichkeit zu skalieren. Variante C ist in diesem Szenario also ebenfalls eine Mischform von ÜK und HM, jedoch liegt das Gewicht stärker bei der HM als in Variante B.

29.4 Resultate

Sämtliche 10 Karten der Variante B werden hier wiedergegeben (Abb. 137). Anhand der Einfärbung lässt sich Variante A nachvollziehen (Foren auszählen). Auf die Wiedergabe der Variante-C-Karten wird verzichtet, da sie keinen Informationsgewinn bringen. Anmerkung: Die Einführungstexte der Foren sind mit einer helleren Schattierung und in der Bezeichnung mit einem Unterstrich markiert – das ist an dieser Stelle irrelevant und soll nicht verwirren.

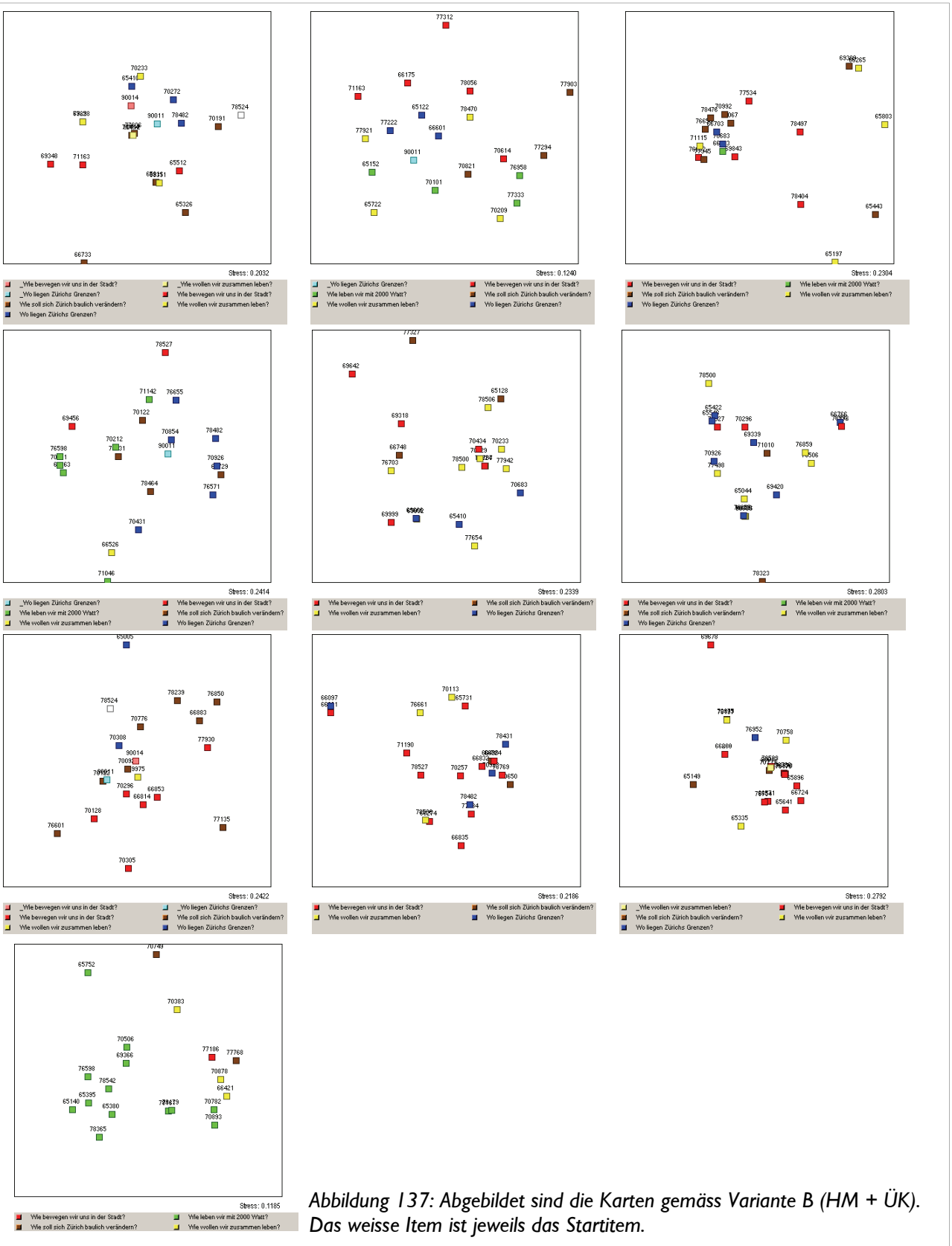


Abbildung 137: Abgebildet sind die Karten gemäss Variante B (HM + ÜK). Das weisse Item ist jeweils das Startitem.

Tabelle 4 gibt einen Überblick dieser verschiedenen Methoden. Die Spaltenüberschriften bedeuten:

- *Forum gepostet:* vom Verfasser des Beitrages ausgewähltes Forum
- *Charakteristik:* Forum, welches nach unseren subjektiven Kriterien passen würde
- *Variante A:* das meistgenannte Forum der 20 Texte (falls mehrere Foren dieselbe maximale Punktzahl aufweisen, sind alle Foren angegeben)
- *Variante B/C:* Forum, welches die Nachbarschaft des Startitems visuell dominiert

Fett gedruckt ist jeweils das Forum in derjenigen Spalte, welche für das betreffende Item die besten Resultate erzielte.

Tabelle 14: Vergleich der verschiedenen Methoden zur Bestimmung des passenden Forums.

ID Startitem	Forum gepostet	Charakteristik	Variante A	Variante B	Variante C
78464	Bauen	Bauen, Zusammenleben	Bauen, Mobilität, Zusammenleben	unklar	**Bauen, Zusammenleben**
78470	Zusammenleben	Zusammenleben	Mobilität	Grenzen, Mobilität	unklar
78476	Bauen	Bauen	Bauen	**Bauen**	Bauen, Mobilität
78482	Grenzen	Grenzen	Grenzen, 2000 Watt	**Grenzen**	**Grenzen**
78500	Zusammenleben	Zusammenleben	**Zusammenleben**	**Zusammenleben**	unklar
78506	Zusammenleben	Zusammenleben	Grenzen	unklar	**Zusammenleben**
78524	Mobilität	Bauen	**Bauen**	**Bauen**, **Grenzen**	unklar
78527	Mobilität	Mobilität	**Mobilität**	**Mobilität**	unklar
78533	Mobilität	Mobilität, Zusammenleben	Mobilität	**Mobilität, Zusammenleben**	Mobilität
78542	2000 Watt	2000 Watt	**2000 Watt**	**2000 Watt**	**2000 Watt**

Das simple Auszählen der ÜK-ähnlichsten Items funktioniert bei vier der 10 Texte. Skaliert man die ÜK-ähnlichsten Texte – Variante B – wird das korrekte Forum sieben Mal eindeutig erkannt, bei zweien ist das Ergebnis nicht eindeutig und bei einem falsch (Item 78470). Allerdings scheint dieses Item für alle Methoden schwierig zu fassen zu sein. Variante C funktioniert etwas weniger gut als Variante B und muss vier Mal passen («unklar»).

Interessant ist vor allem Item 78524: Das gepostete Forum «Mobilität» passte nur entfernt; «Bauen» wäre richtig gewesen. Variante A erkannte dies. Variante B ebenfalls, legte aber auch einige Items des Forum «Grenzen» in dessen Nähe. Tatsächlich geht es im Beitrag auch um Grenzen – um klar definierte Parks:

> Die eine sache ist die Bewegung, Wachstum für alle. Aber was macht die Qualität von Zürich aus? Das grosse Hektik, das viel Verkehr und gleichzeitig die grosse Parkanlagen das See und die viele Plätze. Mein Anliegen ist, dass man ein Platz, ein Park

> nicht architektonisch sinnvoll gestalten, bebauen kann, soll. Bitte kein zweiten Brunnen, neben der alte Brunnen (Hottingen)und das schöne freie Dorfplatz bemöblieren, so dass es kein Märkte mehr stattfinden können. Oder in Zürich Nord wo mit Pflanzen Architektur nachbaut wird und durch die Domestizierung der Pflanzen die Entstehung von etwas neuen verweigert wird. Es ist alles geplant. Stadtentwicklung und dessen viel besprochenes anstrebenswertes Vielfalt und die Freiheit durch nicht definierte Freiräume betrifft Plätze und Parks auch. Die Rahmen müssen klar definiert werden aber bitte Parks durch Gärter gestalten lassen und Plätze nicht durch Architektur Wettbewerbe kreativ verunstalten. Ich brauche nicht immer noch krativer aussehende Plätze sonder brauchbare Plätze die mir zu meiner Erholung beitragen, durch lebendige Schattenwurf, wachsen, veränderungen wie ein Baum zum Beispiel... Schaut die Römerhofplatz Gestaltung sehr hübsch von weitem, aber absolut unbrauchbar. Nur ein Baum anstatt der Häuschen oder was das ist würde mehr bringen. Bitte, baut mehr Bäume!

Testhalber wurde anhand zweier Startbeiträge untersucht, wie die Karten aussehen, wenn nicht erst die 100 ÜK-ähnlichsten Texte gesucht, sondern direkt die 20 Hof-ähnlichsten aus dem Textkorpus gezogen werden.

Für Item 78464 klappt das hervorragend: Das Startitem ist umgeben von Texten der Foren «Bauen» und «Zusammenleben»; für Item 78470 lohnt sich der Rechenaufwand hingegen nicht: Der Text ist in der Tat uneindeutig, aber der Schwerpunkt «Zusammenleben» ist nicht einmal in der Nähe zu finden.

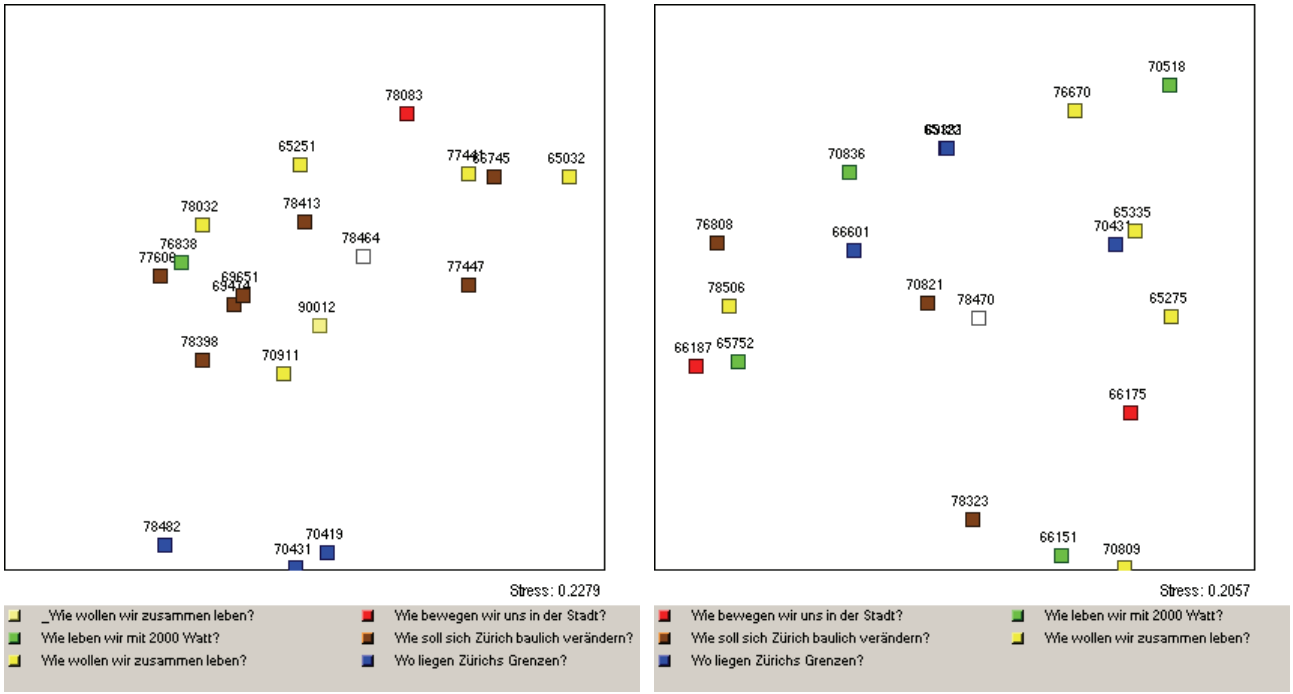


Abbildung 138: Bei beiden Karten wurde die Hof-ähnlichsten Texte aus dem Gesamtdatenbestand herausgesucht.

29.5 Diskussion

Variante B (ÜK + HM) eignet sich hervorragend, um das inhaltlich passende Forum eines neuen, geposteten Beitrages zu verifizieren, beziehungsweise um ein besseres Forum vorzuschlagen. Gerade die

Karten zeigen deutlich, wenn ein Beitrag nicht nur falsch kategorisiert wurde, sondern auch, wenn ein Beitrag inhaltlich zu mehreren Foren passen könnte.

Vergegenwärtigt man sich das Anwendungsszenario, liessen sich solche Mehrfachzuordnungen sogar vermeiden. Da fortwährend die Forumskategorisierung geprüft und dem Verfasser zurückgemeldet würde, kristallisierte sich im zeitlichen Verlauf ein bestimmtes Forum für dieses Thema heraus. Ein Beitrag kann zwar weiterhin inhaltlich in zwei verschiedene Foren passen, das liegt in der Natur der Sache, aber das System unterstützte die Beitragsverfasser in einer einheitlichen Kategorisierung.

Dabei ist zu beachten, dass das gezeigte Verfahren erst arbeiten kann, wenn eine gewisse Textbasis aufgebaut ist. Erst dann können a) die Keywords und b) eine Vergleichsmenge von Texten generiert werden.

Anhang

http://www.geluck.com/

A1 Abkürzungen

API: Application Programming Interface (maschinelle Schnittstelle eines Programmes)
DEM: Dreiecksmatrix (zur Aufnahme von Paarähnlichkeitswerten)
GUI: Graphical User Interface (grafische Benutzeroberfläche eines Programmes)
HM: Hofmethode (Kernalgorithmus zur Berechnung von semantischen Textähnlichkeiten)
HS: Hierarchisches Sortieren (Verfahren zur Ermittlung von Objektähnlichkeiten)
IR: Information Retrieval (Fachgebiet der Informatik: Informationsrückgewinnung)
KWII: KeywordII-Verfahren (Verfahren zur Bestimmung von Keywords, s. Kap. 15)
MA: Mitarbeiter/Mitarbeiterin
NMDS: Nonmetrische Multidimensionale Skalierung (Verfahren, um (Un-)Ähnlichkeiten von Objekten in einer niedrigdimensionalen Karte abzubilden)
PS: Paralleles Sortieren (Verfahren zur Ermittlung von Objektähnlichkeiten)
RVK-Liste: Keywords, die aus dem Katalog der «Regensburger Verbundklassifikation» stammen
SM: SemanticMapper (von uns programmiertes Framework, in dem die HM und die verschiedenen Hilfsalgorithmen implementiert sind, zudem Umgebung zur Erhebung von Experimenten und Auswertungen)
ÜK: Überlappungskoeffizient (zur Berechnung von Listenähnlichkeiten)
VP(n): Versuchsperson(-en)
WF: Wortfrequenzmethode (Verfahren zur Bestimmung von Keywords, s. Kap. 14)

A2 Ergänzendes Material

A2.1 Beschreibung des Webforums Stadtdebatte

Die Stadt Zürich führte im September 2011 eine Online-Stadtdebatte[35] durch. Ziel war den Informationsaustausch mit der Bevölkerung zu fördern und neue Ideen für ein «gemeinsames» Zürich zu generieren. Während dreier Tagen konnte sich die Bevölkerung zu verschiedenen Themen äussern. Die Debatte war offen: Jede interessierte Person konnte daran teilnehmen, sofern sie sich im Vorfeld registriert hatte.

Abbildung 139: Empfangsseite der Stadtdebatte

35 http://www.stadt-zuerich.ch/stadtdebatte

Umgesetzt wurde die Debatte als moderiertes Webforum in Zusammenarbeit mit IBM. Als technische Umgebung diente die Jam-Software[36] von IBM. Jam ist gleichzeitig der Name der Software, wie auch die Bezeichnung für diese Art von kollaborativem Generieren von Ideen und Diskutieren mit Moderation. Grundsätzlich ist es ein Forum: Jemand erstellt ein neues Thema (Thread), andere können dazu Stellung nehmen.

Im Falle der Stadtdebatte gab es fünf Foren, siehe auch Screenshot oben. Die folgende Liste zeigt die Farbcodierung, wie sie die in dieser Arbeit verwendet wird. In Klammern stehen die Begriffe, mit denen die betreffenden Foren abgekürzt werden, sowie die Anzahl der darin geposteten Beiträge:

- Wie soll sich Zürich baulich verändern? (Bauen, 492)
- Wo liegen Zürichs Grenzen? (Grenzen, 219)
- Wie wollen wir zusammen leben? (Zusammenleben, 375)
- Wie leben wir mit 2000 Watt? (2000 Watt, 250)
- Wie bewegen wir uns in der Stadt? (Mobilität, 660)

Jedes Forum wurde mit einem Einleitungstext beschrieben. Darunter waren die einzelnen Threads aufgelistet, innerhalb deren die eigentlichen Diskussionen stattfanden. Jedes Forum wurde von einem Moderator betreut (O. M. war einer davon). Dieser hatte die Aufgabe, bei Bedarf unterstützend

Abbildung 140: Beispiel Forum Bauen: Einleitungstext und die einzelnen Threads (Screenshot zur besseren Übersicht bearbeitet)

einzugreifen; beispw. nachfragen, wenn etwas unklar war; konkrete Beispiele fordern, wo jemand zu pauschal war; auf andere Foren hinweisen, wo das

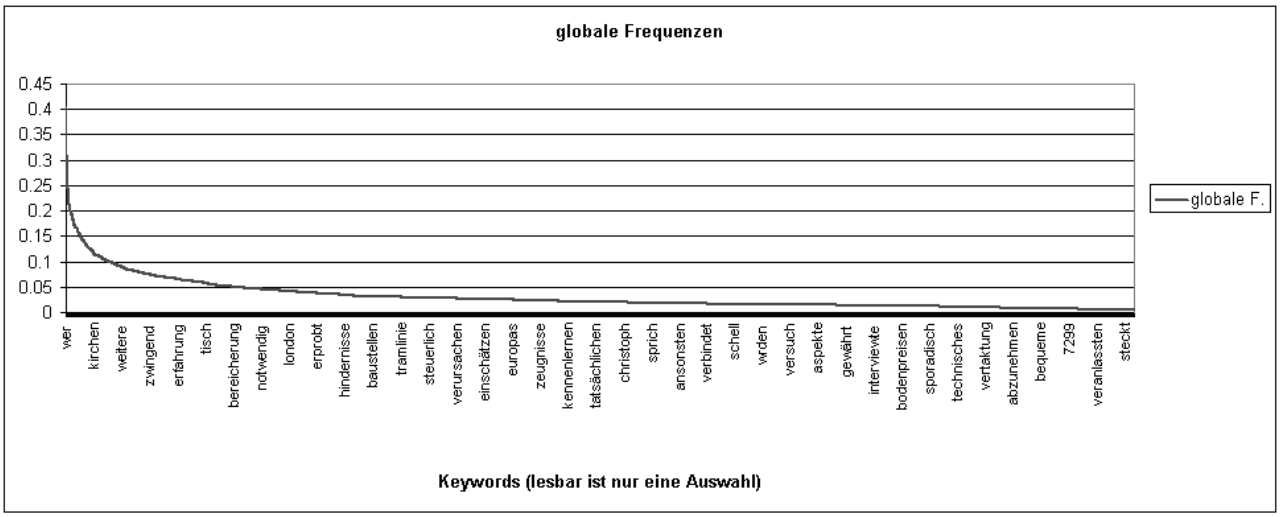
Abbildung 141: Beispiel einer Diskussion (Thread)

36 https://www.collaborationjam.com/

Thema besser passen würde; interessante Beiträge hervorheben, beziehungsweise prominent platzieren. Der Moderator hätte auch Beiträge löschen müssen, wenn diese Werbung oder Beleidigungen enthalten hätten. Soweit mir bekannt ist, war das während der gesamten Debatte nicht nötig (!).

Die Debatte war ein Erfolg. Die Leute zeigten sich engagiert und es kamen 2'000 Beiträge zusammen, mehrheitlich auf einem hohen Niveau. Rund 200 User waren aktiv, d.h. schrieben Beiträge.

A2.2 Anhang zu Kapitel 14, Wortfrequenzmethode: Auswahl der Keywords mittels Überlappungskoeffizient



Anhang 142: Verteilung der Wortfrequenzen, 190 Wörter, berechnet nach der Wortfrequenzmethode, globale Textbasis (Kap. 14)

Anhang 11: Die gesamte Wortliste der Wortfrequenzmethode, nachdem sie manuell bereinigt wurde. (Kap. 14)

2000
altstetten
angemessenen
arbeite
arbeitseinkommen
arbeitsplätze
auto
bahn
bau
bedürfnisse
befürworten
beipflichten
beispiel
bereits
bezahlen
bleiben
braucht
bruno
bucheli
bürger
bus
charme
coole
danke
denken
diskussionsteilnehmende
ebene
eignen
einfach
einwohner
einzigartig
energien
entwicklung
erhöhen
ersetzen
fahren
fahrplangestaltung
fall
finden
fläche
fördern
frage
freie
freiräumen
fussgänger
gegenüber
gehen
gemeinden
genau
genutzt
geplant
gerade
gesellschaft
gesprochen
grenzen
grünflächen
grünwald
gruss
gut
gute
guten
hagelt
hallo
halten
hardbrücke
hecken
heimat
heimatgefühle
herr
herzlich
heute
hoch
hochhäuser
höher
idee
individualverkehr
infrastruktur
innenstadt
jahren
jemand
kalkbreite
kanton
kasernenareal
kennt
kirchen
kommen
kommt
kommunaler
kongresshaus
konkrete
kosten
kreis
kreuzungen
krippenplatz
kurze
lebensqualität
leisten
leute
liebe
lösen
lugano
luzern
mal
mehr
meinung
meter
mindestens
minute
mobil
möglich
monetärer
nähe
nationaler
nehmen
nötig
öffentlichen
ort
paar
park
personengruppen
platz
postneoliberalismus
probieren
profit
projekt
proteste
prozess
quartiere
recht
rechts
reduzieren
region
regionaler
rekurse
rund
rütihof
schneller
schön
sehen
sollen
stadt
stehen
stellen
steuern
strassen
teilweise
thema
tower
trams
trauben
tun
überall
umgesetzt
umlandgemeinden
unterwegs
vbz
velofahrerinnen
velos
verbesern
verbesserungsvorschläge
verfügung
verhältnis
verkehr
vermehrt
verständnis
video
viele
vision
wachsen
wasser
watt
weniger
wenigstens
wert
weshalb
wieso
willkommen
wohl
wohnraum
wohnung
wohnungen
work
wurde
würdest
zeigen
zeit
ziel
zug
zürich
zusammenleben

Anhang 2: Die 16 Textsample, welche im Kapitel über die Wortfrequenzmethode benutzt wurden. (Kap. 14)

ID: 64987 **Forum:** Wie soll sich Zürich baulich verändern?

Titel: vereinfachte baubewilligungen

Beitrag:
wer in zürich baut oder bauen will, muss sich durch ein dschungel von vorschriften kämpfen. manchmal hat man das gefühl, die behörden legen einem lieber steine in den weg anstelle zu helfen.</br>manch privater eigentümer lässt seine liegenschaft lieber im jetzigen zustand als z.b. den brachliegenden dachstock auszubauen.</br></br>mein vorschlag: verbesserte nutzungsvorschriften, vereinfachte und kürzere verfahren und weniger hohe gebühren.

ID: 65950 **Forum:** Wie wollen wir zusammen leben?

Titel: Heterogenität

Beitrag:
Heterogenität in den Quartieren ist mir ein ganz wichtiges Anliegen. Wie lässt sie</br>sich jedoch verwirklichen in Quartieren, in denen wegen der Wohnungspreise fast nur</br>noch überdurchschnittlich Verdienende leben können? Wie lässt sich gegenseitiges</br>aufeinander Zugehen praktizieren in Wohngegenden, in denen die gestressten Berufsleute</br>nur noch abends spät zum Schlafen in ihr Quartier zurückkehren? Zwei Tendenzen, die</br>beträchtlich zunehmen und meiner Meinung nach nicht zur Lebensqualität beitragen.

ID: 66178 **Forum:** Wie wollen wir zusammen leben?

Titel: 24-Stunden-Stadt?

Beitrag:
Die 24-Stunden-Stadt ist in aller Munde. Aufgabe der Stadt ist es, ein Gleichgewicht zwischen lebendiger Urbanität und den Ruhe-Bedürfnissen der Wohnbevölkerung zu gewährleisten. Nicht ganz einfach: Was für die einen eine grosse Bereicherung und mehr Leben darstellt, bedeutet für die anderen vor allem mehr Lärm und Abfall. </br></br>Wie erleben Sie Zürich? Festhütte (WM, EM, Feste im Quartier, etc.) oder Oase der Ruhe? </br>Wie viel Party ist Ihrer Meinung nach angemessen?

ID: 69543 **Forum:** Wie bewegen wir uns in der Stadt?

Titel: Fahrrad als must

Beitrag:
Ich habe 5 Autos und ebenfalls einen solchen Sportwagen, aber einen schnelleren und älteren als Mark's. An die Arbeit pendle ich aus Zollikon mit meinem Velo (davon habe ich 16)und zwar Sommer und Winter. An ca. 6 Tagen im Jahr muss ich den ÖV nehmen, weils zu stark regnet. Wenn ich das Auto nehme für die Stadt und zwischen all den genervten Cayenne Fahrern stehe, bin ich selbst schuld und rege mich nicht auf.</br></br>Mein Konzept: Autopendler bleiben draussen und wir versuchen die Städteintiative in Zürich zu realisieren - ein Konzept für den Langsamverkehr mit vielen Velospuren so ein bisschen wie in Kopenhagen. Cruisen wird wieder möglich, wie schön für die Stadt! </br>Mein Konzept: Autopendler bleiben draussen und wir versuchen die Städteintiative in Zürich zu realisieren - ein Konzept für den Langsamverkehr mit vielen Velospuren so ein bisschen wie in Kopenhagen. Cruisen wird wieder möglich und ich kann wieder mal mit dem Ferrari in die Stadt ohne Olio caldo.

ID: 69708 **Forum:** Wie soll sich Zürich baulich verändern?

Titel: Wohnen und arbeiten in Zürich: okologisch sinnvoll

Beitrag:
In Zürich verdichtet und damit in die Höhe zu bauen, ist nur richtig sondern wurde bereits so beschlossen. Jetzt geht es an die Umsetzung!</br></br>Sinnvoll ist es, wenn Zürich weiter wachsen kann. So, dass es genügend Wohnraum hat - vor allem auch für diejenigen, die in der Stadt arbeiten. Denn wer in der gleichen Stadt wohnt und arbeitet, reduziert die Umweltbelastung massiv.

ID: 70239 **Forum:** Wie bewegen wir uns in der Stadt?

Titel: Langsamverkehrsweg nicht nur für Velos

Beitrag:
Ein durchgehendes Verkehrsnetz für den Langsamverkehr müsste so beschaffen sein, dass der Platz für alle, die sich dazu zählen, reicht. ein Inlineskater sollte einen elektrischen Rollstuhl überholen können und der E-Bike-Trailer von Sihlcity an einer Familie mit Velo, Kinderanhänger + Likeabike vorbeikommen. Und der Platz müsste reichen, dass ältere Menschen frühzeitig auf ihr Auto verzichten und auf ein Elektromobil umsteigen und so bis ins hohe Alter mobil bleiben könnten. Und er müsste so sicher sein, dass Kinder und Jugendliche selbständig (mit Velo, Kickboard, Skateboard) unterwegs sein könnten und es für sie mega-peinlich wäre, von Mami gefahren zu werden. Das würde soviele Probleme unserer Gesellschaft lösen. Wieso tut sich Zürich so schwer damit? Wie wäre es, wenn man Jan Gehl aus Kopenhagen als Berater beiziehen würde http://de.wikipedia.org/wiki/Jan_Gehl.

ID: 77126 **Forum:** Wie soll sich Zürich baulich verändern?

Titel: Verdichten = sparen

Beitrag:
Verdichten an einem Ort, damit an einem anderen mehr freier Raum entsteht. Wenn man verdichtet baut geht man sparsam mit dem Raum um, verbraucht weniger Ressourcen pro Person, weniger (und effizientere) Infrastruktur, kürzere Wege und lässt mehr Raum für anderes. Wenn man nicht verdichtet baut, dann nehmen die Transportwege zu, die Strassen nehmen zu, die Wasserleitungen, Abwasserleitungen, Stromleitungen usw. und damit werden viel mehr Flächen zubetoniert.

ID: 77780 **Forum:** Wie bewegen wir uns in der Stadt?

Titel: Slow Town: tiefe, gleichmässige Geschwindigkeiten

Beitrag:
Es ist eigentlich ganz einfach: Tiefe Geschwindigkeiten (Tempo 40 auch auf Kantonsstrassen gesetzlich möglich) zur Ko-Existenz aller Strassenbenützer! Es geht nicht, dass einzelne (Auto-)Raser die Wohnquartiere als Spielplatz benützen, um den Ladies zu imponieren. 60 oder auch mal 70 km/h sind garantiert tödlich. Tödlich auch für jegliches Zusammen-Leben im Quartier.

A2.3 Anhang zu Kapitel 16, Tagcloud-Verfahren von Semager

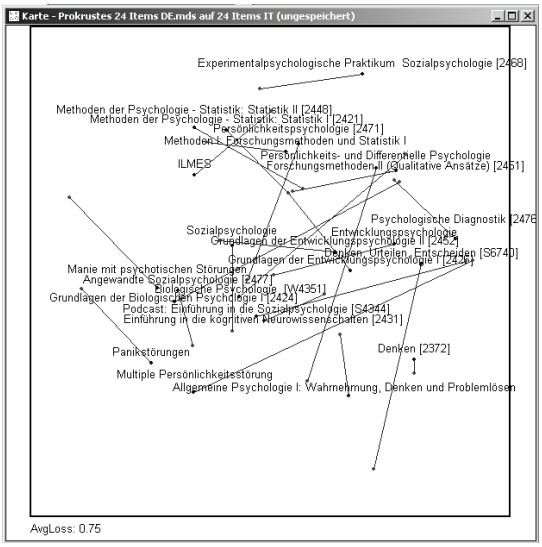
Anhang 3: Es hat keinen grossen Einfluss, ob dem Tagcloud-API die entrauschten Texte (linke Spalte) oder die Originaltexte (rechte Spalte) übergeben werden. Die unterschiedlichen Wörter sind fett markiert.

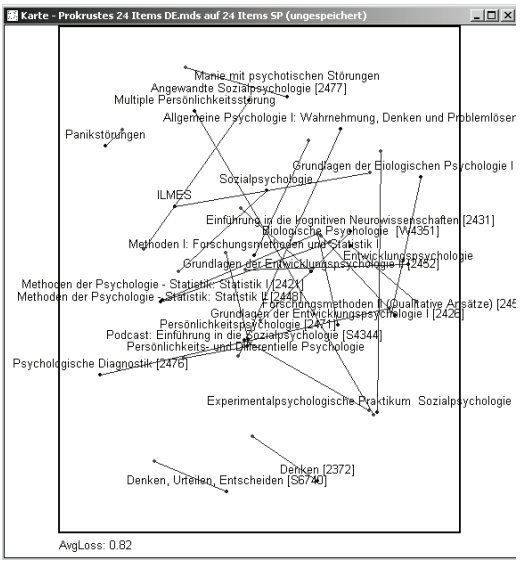
text_texted	full_text
affektive	affektive

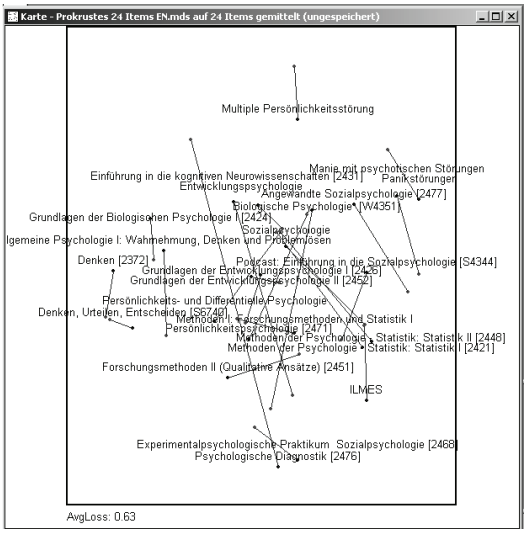
affektiven	affektiven
analyse	analyse
angst	angst
anhand	anhand
ansätze	ansätze
auftreten	auftreten
ausgewählte	ausgewählte
behandelt	behandelt
betroffene	betroffene
betroffenen	betroffenen
darstellung	darstellung
daten	daten
denkens	denkens
devil	devil
einführung	einführung
	empirischen
entwicklung	entwicklung
entwicklungspsychologie	entwicklungspsychologie
episode	episode
episoden	episoden
ergebnisse	ergebnisse
experimentalpsychologische	experimentalpsychologische
experimente	experimente
experimentellen	experimentellen
expra	expra
forschung	
forschungsmethoden	forschungsmethoden
frage	frage
	gebrauch
gegenwärtig	gegenwärtig
grundlagen	grundlagen
gruppen	gruppen
insomnie	insomnie
kinder	kinder
	konsequenzen
kriterien	kriterien
leistung	**lernzentrum**
mesosworld	mesosworld
	methoden
motivation	motivation
nichtparametrische	nichtparametrische
personen	personen
podcast	podcast
praktikum	praktikum
probleme	probleme
psychologie	psychologie
psychotische	psychotische
schizoaffektive	schizoaffektive
schizophrenie	schizophrenie
schwere	schwere

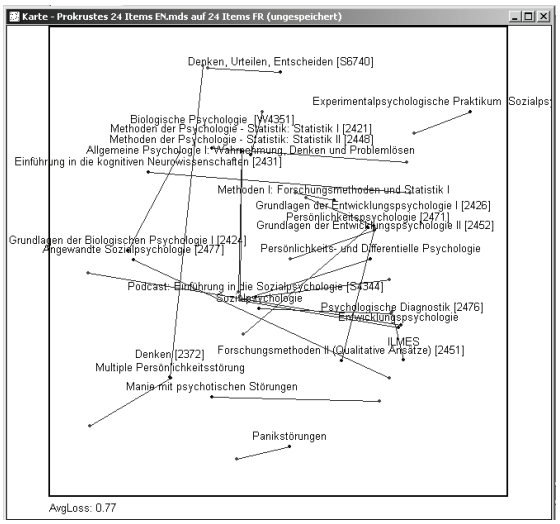
	soziale
sozialpsychologie	sozialpsychologie
statistischen	statistischen
stimmung	stimmung
störung	störung
störungen	
studierenden	studierenden
stunden	
symptome	symptome
test	test
verfahren	verfahren
	verhalten
vorlesung	vorlesung
wahrnehmung	wahrnehmung

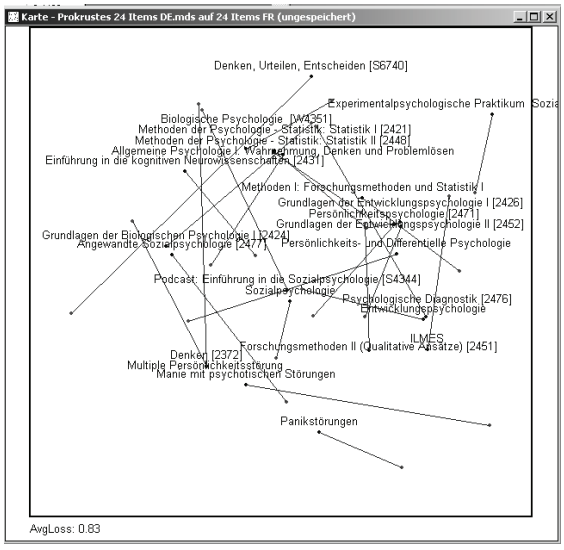
A2.4 Anhang zu Kapitel 21, Multilinguality I

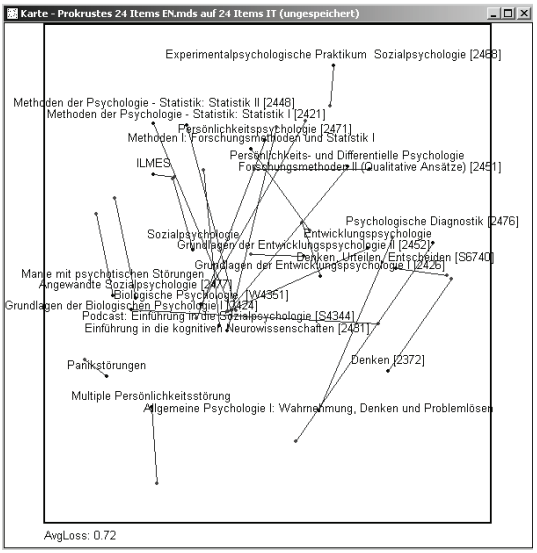


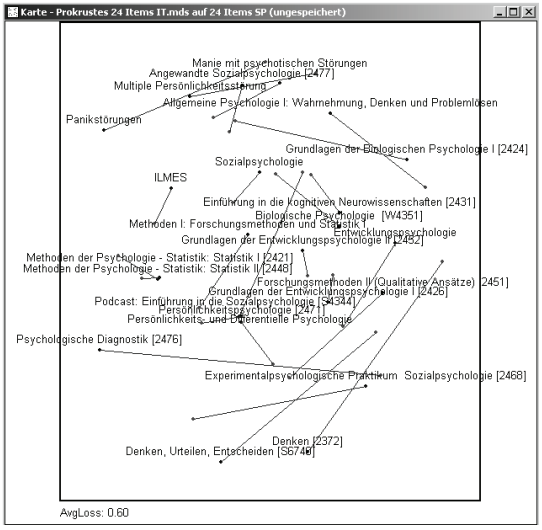


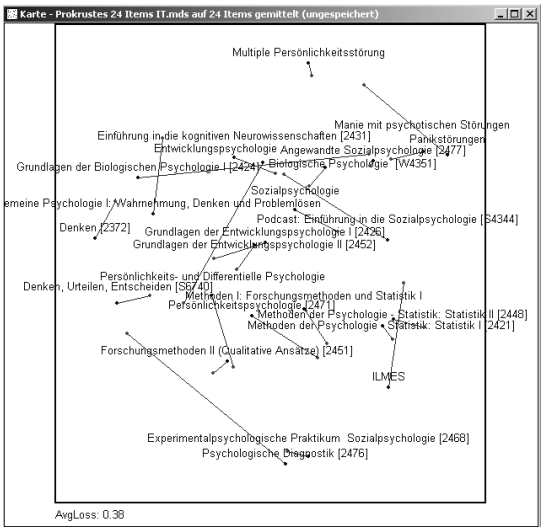


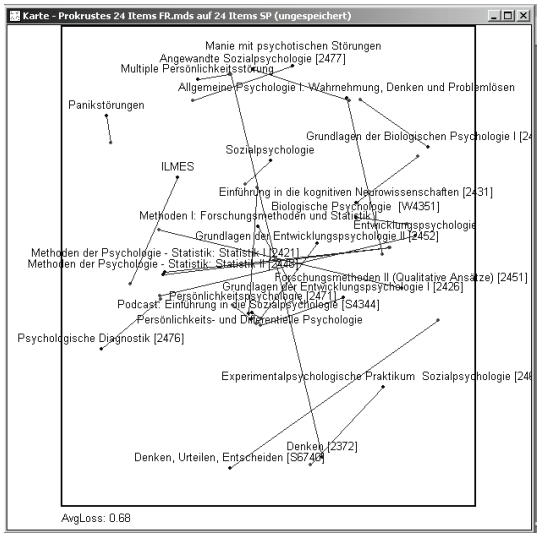


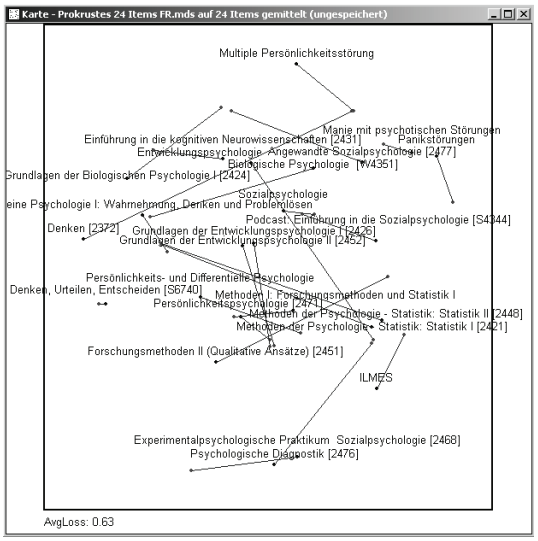

Karte - Prokrustes 24 Items DE.mds auf 24 Items FR (ungespeichert)
Denken, Urteilen, Entscheiden [S6740]
Experimentalpsychologische Praktikum
Biologische Psychologie [W4351]
Methoden der Psychologie - Statistik: Statistik I [2421]
Methoden der Psychologie - Statistik: Statistik II [2448]
Allgemeine Psychologie I: Wahrnehmung, Denken und Problemlösen
Einführung in die kognitiven Neurowissenschaften [2431]
Methoden I: Forschungsmethoden und Statistik I
Grundlagen der Entwicklungspsychologie I [2426]
Persönlichkeitspsychologie [2471]
Grundlagen der Entwicklungspsychologie II [2452]
Grundlagen der Biologischen Psychologie I [2424]
Angewandte Sozialpsychologie [2477]
Persönlichkeits- und Differentielle Psychologie
Podcast: Einführung in die Sozialpsychologie [S4344]
Sozialpsychologie
Psychologische Diagnostik [2476]
Entwicklungspsychologie
ILMES
Forschungsmethoden II (Qualitative Ansätze) [2451]
Denken [2372]
Multiple Persönlichkeitsstörung
Manie mit psychotischen Störungen
Panikstörungen
AvgLoss: 0.83

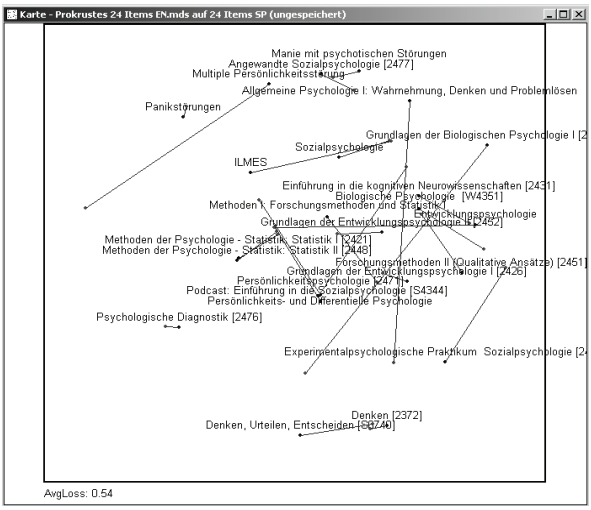

Karte - Prokrustes 24 Items EN.mds auf 24 Items IT (ungespeichert)
Experimentalpsychologische Praktikum
Sozialpsychologie [2468]
Methoden der Psychologie - Statistik: Statistik II [2448]
Methoden der Psychologie - Statistik: Statistik I [2421]
Persönlichkeitspsychologie [2471]
Methoden I: Forschungsmethoden und Statistik I
ILMES
Persönlichkeits- und Differentielle Psychologie
Forschungsmethoden II (Qualitative Ansätze) [2451]
Psychologische Diagnostik [2476]
Sozialpsychologie
Entwicklungspsychologie
Grundlagen der Entwicklungspsychologie II [2452]
Denken, Urteilen, Entscheiden [S6740]
Grundlagen der Entwicklungspsychologie I [2426]
Manie mit psychotischen Störungen
Angewandte Sozialpsychologie [2477]
Biologische Psychologie [W4351]
Grundlagen der Biologischen Psychologie I [2424]
Podcast: Einführung in die Sozialpsychologie [S4344]
Einführung in die kognitiven Neurowissenschaften [2431]
Panikstörungen
Denken [2372]
Multiple Persönlichkeitsstörung
Allgemeine Psychologie I: Wahrnehmung, Denken und Problemlösen
AvgLoss: 0.72

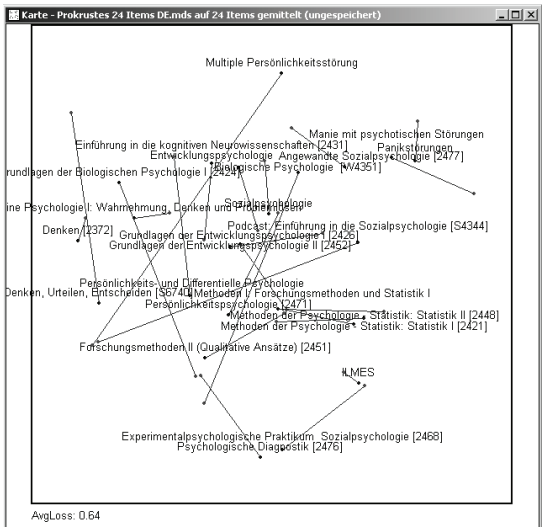

Karte - Prokrustes 24 Items IT.mds auf 24 Items SP (ungespeichert)
Manie mit psychotischen Störungen
Angewandte Sozialpsychologie [2477]
Multiple Persönlichkeitsstörung
Allgemeine Psychologie I: Wahrnehmung, Denken und Problemlösen
Panikstörungen
Grundlagen der Biologischen Psychologie I [2424]
Sozialpsychologie
ILMES
Einführung in die kognitiven Neurowissenschaften [2431]
Biologische Psychologie [W4351]
Methoden I: Forschungsmethoden und Statistik I
Entwicklungspsychologie
Grundlagen der Entwicklungspsychologie II [2452]
Methoden der Psychologie - Statistik: Statistik I [2421]
Methoden der Psychologie - Statistik: Statistik II [2448]
Forschungsmethoden II (Qualitative Ansätze) [2451]
Grundlagen der Entwicklungspsychologie I [2426]
Podcast: Einführung in die Sozialpsychologie [S4344]
Persönlichkeitspsychologie [2471]
Persönlichkeits- und Differentielle Psychologie
Psychologische Diagnostik [2476]
Experimentalpsychologische Praktikum
Sozialpsychologie [2468]
Denken [2372]
Denken, Urteilen, Entscheiden [S6740]
AvgLoss: 0.60

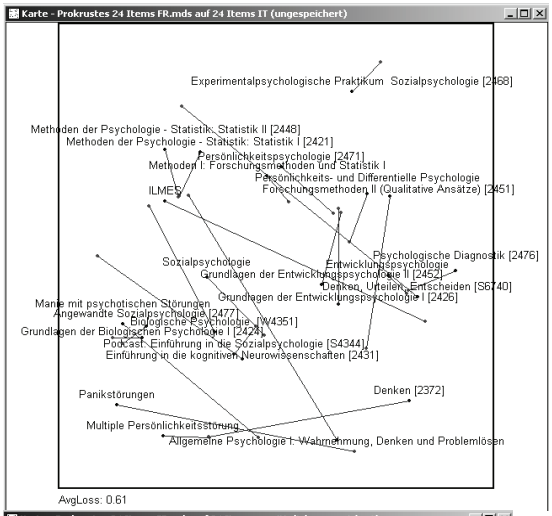

Karte - Prokrustes 24 Items IT.mds auf 24 Items gemittelt (ungespeichert)
Multiple Persönlichkeitsstörung
Manie mit psychotischen Störungen
Einführung in die kognitiven Neurowissenschaften [2431]
Panikstörungen
Entwicklungspsychologie
Angewandte Sozialpsychologie [2477]
Grundlagen der Biologischen Psychologie I [2424]
Biologische Psychologie [W4351]
Sozialpsychologie
Allgemeine Psychologie I: Wahrnehmung, Denken und Problemlösen
Podcast: Einführung in die Sozialpsychologie [S4344]
Denken [2372]
Grundlagen der Entwicklungspsychologie I [2426]
Grundlagen der Entwicklungspsychologie II [2452]
Persönlichkeits- und Differentielle Psychologie
Denken, Urteilen, Entscheiden [S6740]
Methoden I: Forschungsmethoden und Statistik I
Persönlichkeitspsychologie [2471]
Methoden der Psychologie - Statistik: Statistik II [2448]
Methoden der Psychologie - Statistik: Statistik I [2421]
Forschungsmethoden II (Qualitative Ansätze) [2451]
ILMES
Experimentalpsychologische Praktikum
Sozialpsychologie [2468]
Psychologische Diagnostik [2476]
AvgLoss: 0.38

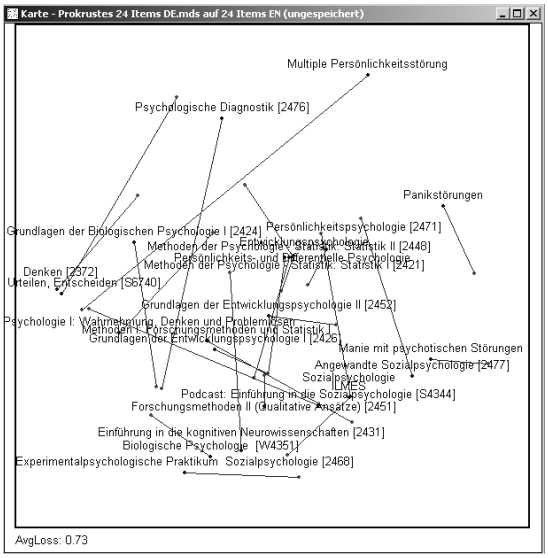

Karte - Prokrustes 24 Items FR.mds auf 24 Items SP (ungespeichert)
Manie mit psychotischen Störungen
Angewandte Sozialpsychologie [2477]
Multiple Persönlichkeitsstörung
Allgemeine Psychologie I: Wahrnehmung, Denken und Problemlösen
Panikstörungen
Grundlagen der Biologischen Psychologie I [2
Sozialpsychologie
ILMES
Einführung in die kognitiven Neurowissenschaften [2431]
Biologische Psychologie [W4351]
Methoden I: Forschungsmethoden und Statistik I
Entwicklungspsychologie
Grundlagen der Entwicklungspsychologie II [2452]
Methoden der Psychologie - Statistik: Statistik I [2421]
Methoden der Psychologie - Statistik: Statistik II [2448]
Forschungsmethoden II (Qualitative Ansätze) [2451]
Grundlagen der Entwicklungspsychologie I [2426]
Persönlichkeitspsychologie [2471]
Podcast: Einführung in die Sozialpsychologie [S4344]
Persönlichkeits- und Differentielle Psychologie
Psychologische Diagnostik [2476]
Experimentalpsychologische Praktikum
Sozialpsychologie [24
Denken [2372]
Denken, Urteilen, Entscheiden [S6740]
AvgLoss: 0.68

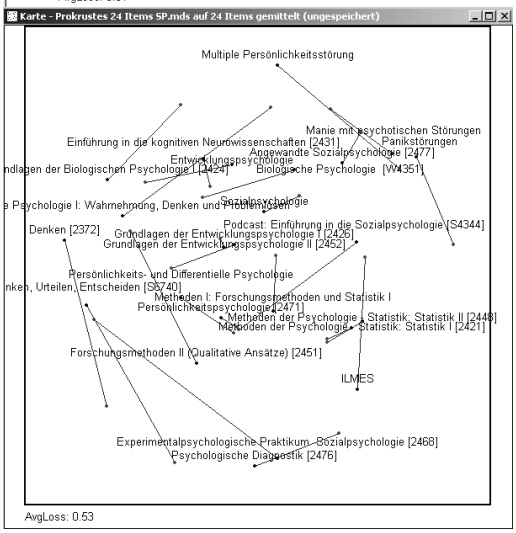

Karte - Prokrustes 24 Items FR.mds auf 24 Items gemittelt (ungespeichert)
Multiple Persönlichkeitsstörung
Manie mit psychotischen Störungen
Einführung in die kognitiven Neurowissenschaften [2431]
Panikstörungen
Entwicklungspsychologie
Angewandte Sozialpsychologie [2477]
Biologische Psychologie [W4351]
Grundlagen der Biologischen Psychologie I [2424]
Sozialpsychologie
eine Psychologie I: Wahrnehmung, Denken und Problemlösen
Podcast: Einführung in die Sozialpsychologie [S4344]
Denken [2372]
Grundlagen der Entwicklungspsychologie I [2426]
Grundlagen der Entwicklungspsychologie II [2452]
Persönlichkeits- und Differentielle Psychologie
Denken, Urteilen, Entscheiden [S6740]
Methoden I: Forschungsmethoden und Statistik I
Persönlichkeitspsychologie [2471]
Methoden der Psychologie - Statistik: Statistik II [2448]
Methoden der Psychologie - Statistik: Statistik I [2421]
Forschungsmethoden II (Qualitative Ansätze) [2451]
ILMES
Experimentalpsychologische Praktikum
Sozialpsychologie [2468]
Psychologische Diagnostik [2476]
AvgLoss: 0.63











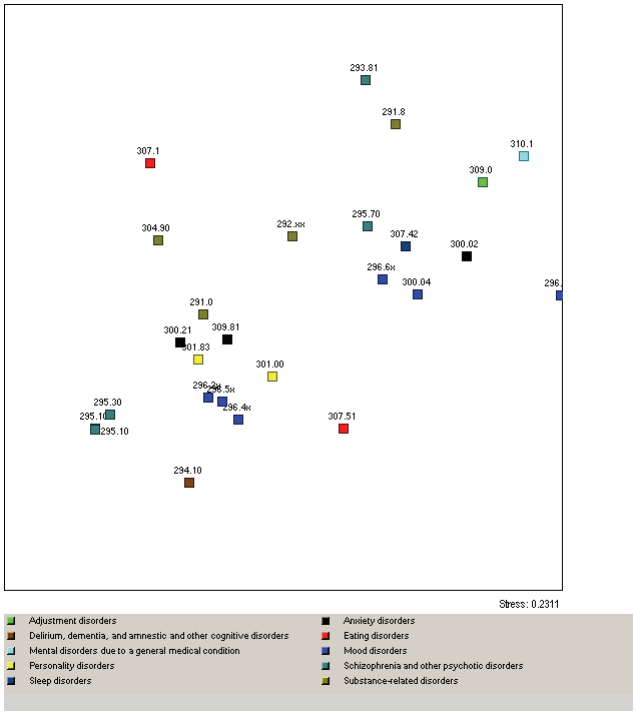
Anhang 4: Prokrustes-Transformation der diverses Sprachkarten: Massive Veränderungen sind zwischen den Sprachen sichtbar, dennoch wird keine Karte eindeutig von den VP präferiert. (Kap. 21)

A2.5 Anhang zu Kapitel 24, Wikipedia-Experiment

Anhang 5: Suchfragen und deren Lösungen

Fragen-Nr.	Text
0	Generelle Fragen (1 Antwort genügt) Als Höhepunkt des Jahres möchtest du mit der Klasse eine Reise unternehmen und einen Erlebnispark besuchen, der sich mit deiner Thematik beschäftigt. Finde solch einen Park! - 20016 Terra Natura (Erlebnispark), oder - 20051 Biosphäre Potsdam
1	Welcher Beruf beschäftigt sich mit dem Thema? - 19721 Umweltingenieurwissenschaften
2	Wie heisst die chinesische, daoistische Theorie zur Naturbeschreibung? - 19572 Fünf-Elemente-Lehre
3	Die besten Arbeiten der Schüler sollen für einen Wettbewerb eingereicht werden. Welcher Anlass eignet sich hierfür? - 19494 Energy Globe Award
4	Wasserspezifische Fragen Wasser hat weitreichende Konsequenzen auf das Wetter. Wie nennt man das Klima, das typischerweise in der Nähe von grossen Gewässern herrscht? - 20035 Seeklima
5	Finde ein typisches Beispiel für einen feuchten Lebensraum (abgesehen von Seen oder Meeren). - 19504 Regenmoor
6	Luftspezifische Fragen Was für eine Rolle spielt die Meteorologie in der Luftfahrt? - 19545 Meteorologie in der Luftfahrt
7	Wie kann man die Raumluft reinigen? - 20056 Ionisator
8	Viele (3 aus vielen) Finde 3 Texte, die sich in philosophischer Weise mit den Themen beschäftigen. - 19473 Vier-Elemente-Lehre - 19482 Archelaos - 19501 Quintessenz (Philosophie) - 19572 Fünf-Elemente-Lehre - 19666 Vaisheshika - 19682 Humoralpathologie - 19873 Seelenmodell - 19986 Timaios (Platon) - 20039 Vastu - 20040 Elisabethanisches Weltbild
9	Finde 3 Beiträge zum Thema Umweltschutz oder auch der Überwachung der Umwelt. - 19478 CO2-Bilanz - 19494 Energy Globe Award - 19513 Mutterboden - 19603 Umweltmonitoring - 19850 Umweltinformationssystem - 19704 Umweltgefährliche Stoffe - 19721 Umweltingenieurwissenschaften - 19778 Umweltverträglichkeit - 19806 Bodenschutzrecht - 19839 Umweltverschmutzung - 19622 Umweltanalytik

A2.6 Anhang zu Kapitel 27, Experiment DSM-IV und Diagnostik



Anhang 6: Die DSM-IV-Texte nach Berechnung der Ähnlichkeitswerte durch den ÜK: Die semantische Strukturierung ist schwach.

A2.7 Weiteres Material

Universität Zürich

2007 Zürich
Kongress SGP
Congrès SSP

The Congress Map: Documenting the similarity of conference contributions by information retrieval from the scientific abstracts

Michel, O., Läge, D.
Universität Zürich - Psychologisches Institut - Angewandte Kognitionspsychologie
(o.michel@psychologie.uzh.ch)

Introduction

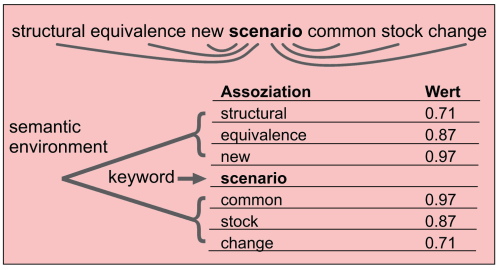
The similarity relation of a number of texts is important not only for congress organisers (who need to group the proposed contributions to meaningful sessions) but to everybody who wants to find certain information within a larger number of texts. Hence, information retrieval methods have been developed to compare texts according to their similarity. We applied four of those methods to model the semantic structure of all the 350 SSP congress contributions. The coefficients of pairwise similarity which derived from these methods are transformed into maps (using a robust version of Nonmetric Multidimensional Scaling).

The four compared methods

These methods are:

- Trigramming: treats text as an undistinguished stream of 3-letter packages
- Überlappungskoeffizient: counts identical words in pairs of texts
- Keyword method: counts the joint occurance of a list of nouns*
- Hofmethode: compares the surrounding semantic environment of a list of nouns*

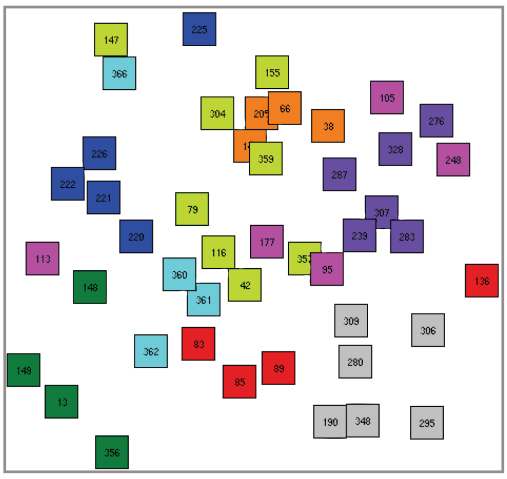
* The list was compiled from all the nouns occuring in the abstract titles.



The Hofmethode takes the surrounding semantic environment of a given keyword and compares this "halo" with the halo of the same keyword in another text. Hence, the meaning of used terms can be compared.

Results

Trigramming and the Überlappungskoeffizient resulted in not very "tidy" maps, still ordering though. The keyword method and the Hofmethode both produced very similar maps of high semantic order. Items, which did not happen to be in their cluster were quite often ambiguous, indeed.



One example of a semantic map, showing the abstracts of all contributions of 10 symposia. Similarites were produced by the Hofmethode.

In cases, where an extensive list of keywords is available (almost 800 nouns for the SSP congress), the keyword method is more efficient. The Hofmethode probably gains performance when there are only few keywords available and more noise in the texts.

Discussion

The presented methods are capable of producing ordered semantic maps which could be a great help for congress organisers as well as for visitors/participants, who are interested to have an overview of the topical range.

Literatur:
Hörmann, H. (1976). *Meinen und Verstehen.* Frankfurt am Main: Suhrkamp.
Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling.* Theory and Applications. New York: Springer.
Marx, W. (1976). Die Messung der Assoziativen Bedeutungs-ähnlichkeit. Zeitschrift für Experimentelle und Angewandte Psychologie, 23(1), 62-76.
Forschungsbericht Hofmethode: http://www.allgpsy.unizh.ch/AKZ-Forschungsberichte/articles/34-AKZ.htm

Anhang 7: Präsentiertes Poster des SGP-Kongresses in Zürich 2007

Universität Zürich

I-KNOW

Semantic structuring of conference contributions by the Hofmethode

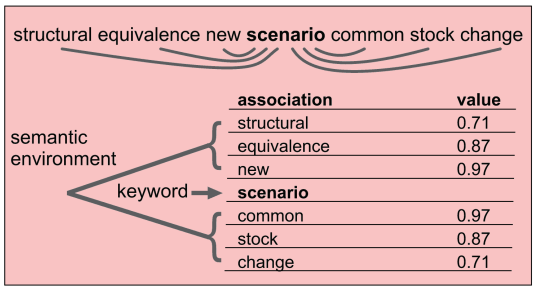
Michel, O., Läge, D.
Universität Zürich - Psychologisches Institut - Angewandte Kognitionspsychologie
(o.michel@psychologie.uzh.ch)

Introduction

The similarity relation of a number of texts is important not only for congress organizers (who need to group the proposed contributions to meaningful sessions) but to everybody who wants to find certain information within a larger number of texts. Because existing information retrieval methods remain mostly on the surface of the words, the resemblance is not primary a semantic one, but a stylistic and vocabulary dependent one. Based on psychological considerations we have developed an algorithm called Hofmethode, which compares the semantic "environment" of key words in order to determine the meaning of the used word in the actual context. Using the example of the SSP congress we show how the Hofmethode can be used to help both congress organizers and participants to find the appropriate contributions.

Functioning of the Hofmethode

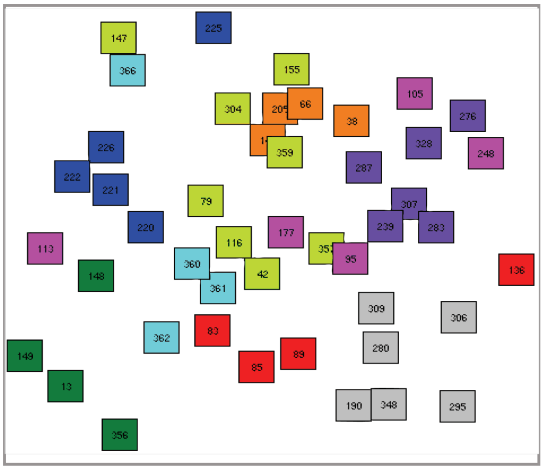
The Hofmethode searches a text for the appearance of a defined set of target words. If found, the semantic environment (halo) is calculated and compared to the halo of the same target word in another text. If similar, than the meaning is regarded to be similar. This procedure was carried out with every target word in 46 abstracts of talks, which were held at the SSP congress, resulting in a triangular matrix of summed up similarity values. These values were transformed into Euclidian distances by means of NMDS and arranged in a semantic map.



The Hofmethode takes the surrounding semantic environment of a given key word and compares this "halo" with the halo of the same keyword in another text. Hence, the meaning of used terms can be compared. The values in the halos correlate to the distance to the target word.

Results

The resulting structure matches very well with the attribution (of the talks to the symposia) done by the congress organizers. The different topics are quite well separated, even though they do not cluster in a narrow sense. Items, which did not happen to be in their cluster, are quite often ambiguous, indeed.



The resulting semantic map of the 46 abstracts. The colours represent the symposium subjects. The overall structure correspondents very well to the attribution of the talks to the symposia by the congress organizers.

Conclusions

Congress organizers could use the Hofmethode as a tool to identify the right contributions for certain congress themes or to form meaningful sessions. Furthermore, visitors of the congress could use the map to gain a fast overview of the available themes. And last but not least: Contributors might find easily other contributions, which are semantically close to theirs.

References:
[Borg, 1997)] Borg, I., & Groenen, P.: Modern Multidimensional Scaling. Theory and Applications. New York: Springer, 1997.
[Hörmann, 1976] Hörmann, H.: Meinen und Verstehen. Frankfurt am Main: Suhrkamp, 1976.
[Marx, 1976] Marx, W.: Die Messung der Assoziativen Bedeutungsähnlichkeit. Zeitschrift für Experimentelle und Angewandte Psychologie, 23(1), 62-76, 1976.
[Michel, 2006] Michel, O. & Läge, D.: Die Hofmethode: Auf dem Weg zum maschinellen Textverständnis (AKZ-Forschungsbericht No. 34). Zürich: Angewandte Kognitionspsychologie, 2006.
[Nohr, 2000] Nohr, H.: Automatische Dokumentindexierung. Fachhochschule Stuttgart, Stuttgart, 2000
[SSP, 2007] SSP-Kongress 2007, 2007, http://www.ssp-sgp2007.ch/
[Wittgenstein, 1960] Wittgenstein, L.: Philosophische Untersuchungen. Unpublished manuscript, Frankfurt, 1960.

Anhang 8: Präsentiertes Poster an der I-KNOW 2010 in Graz.

A3 Zielitems in der NMDS nach Verteilungen bestimmen

In diesem Kapitel wird ein Algorithmus zur Determinierung von Zielitems schrittweise entwickelt, wobei betont werden muss, dass es mehrheitlich Arbeitsideen sind, die nicht weiter ausgeführt werden. Wir haben das Kapitel der Vollständigkeit halber aufgenommen, weil in Kapitel 17 (Repräsentanten-Algorithmus) kurz darauf eingegangen wird und die Arbeitsoberfläche des SemanticMappers Funktionalitäten enthält, die dafür programmiert wurden.

A3.1 Überblick

Der in diesem Kapitel entwickelte Algorithmus sucht aus einer gegebenen NMDS diejenigen Items (Zielitems) heraus, welche die Verteilung in geeigneter Weise widerspiegeln. Diese Zielitems könnten benutzt werden, um weitere Items in die Karte einzupassen, welche nur mit den Zielitems, nicht aber untereinander verglichen werden. Somit wären Kartenberechnungen möglich, die die bisherige – durch Rechenkapazitäten gegebene – Maximalanzahl von ca. 60 Items überschreiten würden.

A3.2 Einleitung

In einer NMDS können prinzipiell zwar beliebig viele Items dargestellt werden, jedoch explodiert die dafür benötigte Rechenzeit mit zunehmender Anzahl und es ist ab ca. 60 Items nicht mehr praktikabel, die Karte in einer Anwendung einem wartendem Benutzer zu präsentieren.

Tabelle 15: Die Rechenzeit für die Erstellung der NMDS steigt mit zunehmender Anzahl Items drastisch an. Erhoben wurden obige Zahlen auf einem PC mit einer Intel Core 2 CPU (2.13GHz), 2GB RAM) und mit ProDaX Version 1.0.7, Standardeinstellungen wurden beibehalten, einzig die Iterationszahl wurde auf von 50 auf 20 variiert (Zahlen in Klammern).

Anzahl Items	benötigte Rechenzeit
20	0.9s (0.4s)
30	1.7s (0.6s)
40	3.1s (2s)
50	7.4s (5.1s)
60	11.8s (10.3s)
80	1min34s (1min30s)
100	2min23s (2min12s)

Die Ähnlichkeitsberechnung der Hofmethode wurde schon stellenweise optimiert und lässt noch Raum für weitere Beschleunigungen. Sinnvoller ist zu diesem Zeitpunkt, die Zeit für die NMDS-Berechnung zu verkürzen.

Hinweis: Im Kapitel 17 (Repräsentanten-Algorithmus) wird ein stark verwandter Algorithmus entwickelt, der aber mehr auf die Dichteverteilung einer Karte Rücksicht nimmt.

A3.2.1 Skizze Lösungsansatz

Angenommen, es sollen 250 bestimmte Texte in der NMDS dargestellt werden. Zuerst werden daraus 30 relevante Dokumente herausgezogen, und zwar mit einem schnellen, etablierten Verfahren, beispw. dem Google-Ranking, einer herkömmlichen Booleschen Suche oder dem Überlappungskoeffizienten. Die 30 Items werden behoft (wie die Zielwörter zustande kommen ist an dieser Stelle noch offen) und eine Karte daraus erzeugt. Der zu erstellende Algorithmus sucht aus dieser Karte diejenigen 15 Items heraus, die die Gesamtstruktur möglichst gut wiedergeben. Künftige Items (beziehungsweise die restlichen 220) werden in die 15er-Karte reingefittet, d.h. sie werden nur mit den 15 Items verglichen, nicht untereinander.

A3.3 Die Algorithmusentwicklung

A3.3.1 Lokale Clusterschwerpunkte

Es steht die Idee des lokalen Clusterschwerpunktes im Fokus: Gibt es in der Karte eine Anhäufung von Items, scheint dort ein bedeutungsähnlicher Schwerpunkt zu sein. Sicherlich sollte ein Item davon als Zielitem ausgewählt werden. Naheliegenderweise nimmt man das zentrale Item. Dieses wird erkoren, indem von jedem Item die Abstände zu allen anderen Items erhoben und aufsummiert wird. Dasjenige Item mit der geringsten Distanzsumme ist das zentrale.

A3.3.2 Clusterschwerpunkte «nur nach Werten»

An einem konkreten Beispiel lässt sich das illustrieren. Anlässlich des SGP-Kongress im September 2007 wurden die eingereichten Abstracts in einer Datenbank gesammelt (s. auch Kap. 22,

Bedeutungsähnlichkeiten von Abstracts: Vier Verarbeitungsebenen). Die Beiträge unterteilten sich in 248 Talks und 103 Posters, insgesamt 351 Beiträge. Arbeiten wir nur mit den Talks, um in die Nähe der 250 Items zu sein. Da es an dieser Stelle um den Algorithmus geht, der die relevanten Items herauszieht, wurden die 30 Items nicht mit einem effektiv praktizierten Verfahren ausgewählt, sondern es wurden die 30 ersten Items der Datenbank entnommen (die ursprüngliche Sortierung der Datenbank beruhte vermutlich auf dem Eingangsdatum der Abstracts und sollte hier keine Rolle spielen).

Abbildung 143 zeigt die dazugehörende NMDS. Die der Karte zugrunde liegende Ähnlichkeiten beruhen auf den summierten Hofähnlichkeiten der Nomina, die in den Abstract-Titeln vorkamen. Eingefärbt ist die summierte Distanz zu allen anderen Items. Je heller das Item, umso kleiner ist die Distanzsumme. Periphere Items sind demnach dunkler.

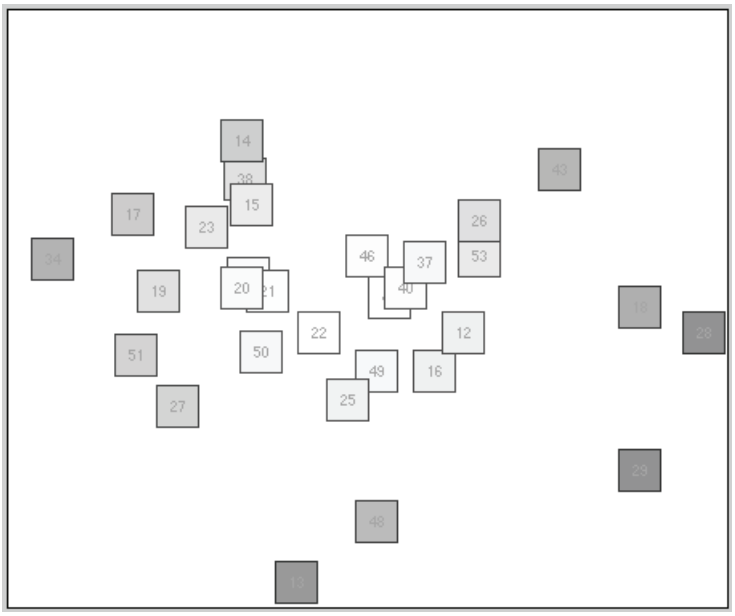


Abbildung 143: Die Items sind nach Nähe zu allen anderen Items eingefärbt: Je heller, umso zentraler.

Färbt man nun der Reihe nach diejenigen Items ein, die die kleinsten Summenwerte besitzen, ergibt sich die Sequenz in Abbildung 144.

Zwar werden mit den ersten zwei Zielitems zwar gleich zwei Cluster identifiziert, danach aber tendiert dieser Algorithmus dazu, die neuen Zielitems in der unmittelbaren Nähe der schon gefundenen Items zu setzen. Das ist inhaltlich nicht erwünscht. Im letzten Bild der Sequenz sieht man deutlich, dass die Peripherie der Karte nicht im geringsten repräsentiert wird.

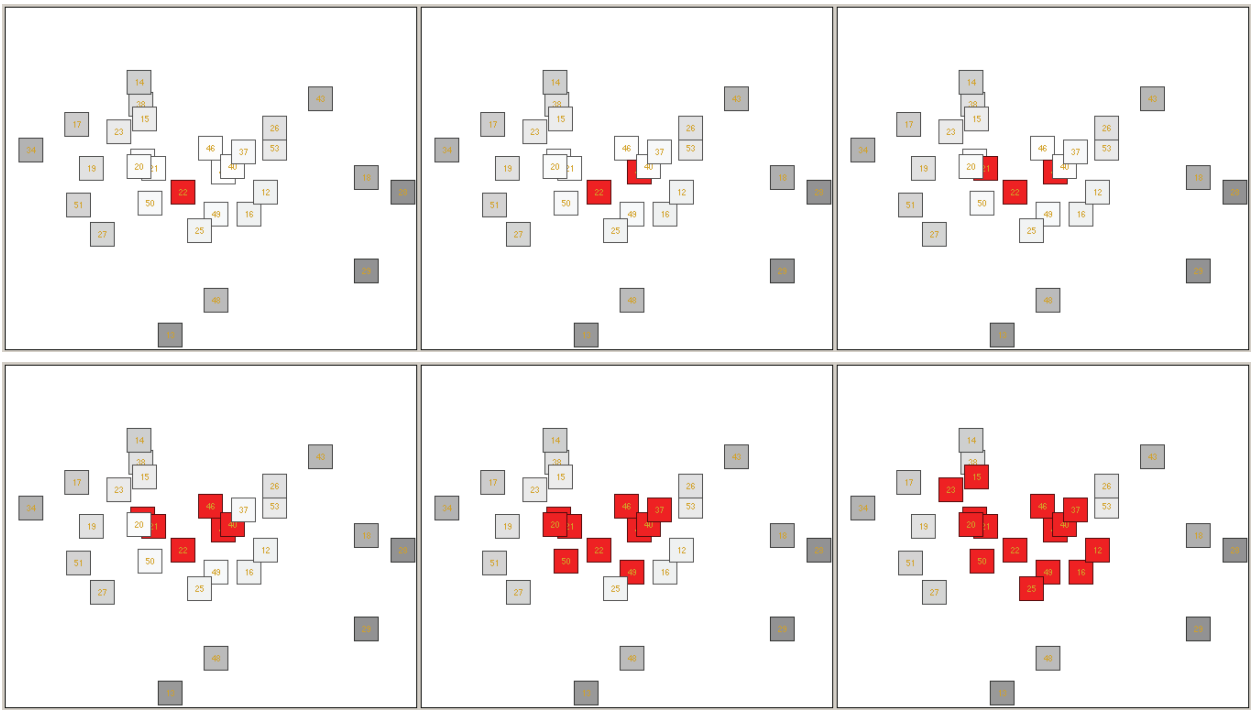


Abbildung 144: Die Sequenz zeigt im zeitlichen Verlauf, wie die Zielitems ausgewählt werden. Die Überrepräsentation der Kartenmitte ist deutlich.

A3.3.3 Clusterschwerpunkte: sichtbarer Horizont

Da das Ziel ist inhaltlich relevante *Bereiche* abzudecken, führen wir den sog. sichtbaren Horizont ein. Die Summendistanz pro Item setzt sich demnach nicht aus den Distanzen zu allen anderen Items zusammen, sondern nur noch zu denjenigen Items, die in der Nähe (innerhalb des sichtbaren Horizonts) liegen. Die Nähe ist parametrierbar und ist ein anzugebender Bruchteil der maximal möglichen Itemdistanz (mD). In Abbildung 145 ist ersichtlich, dass die Items 34 und 28 (ganz links und rechts aussen) die mD ausmachen. Tatsächlich ergibt sich eine etwas bessere Verteilung der Zielitems. Das beste Ergebnis wurde mit 1/10 der mD erzielt.

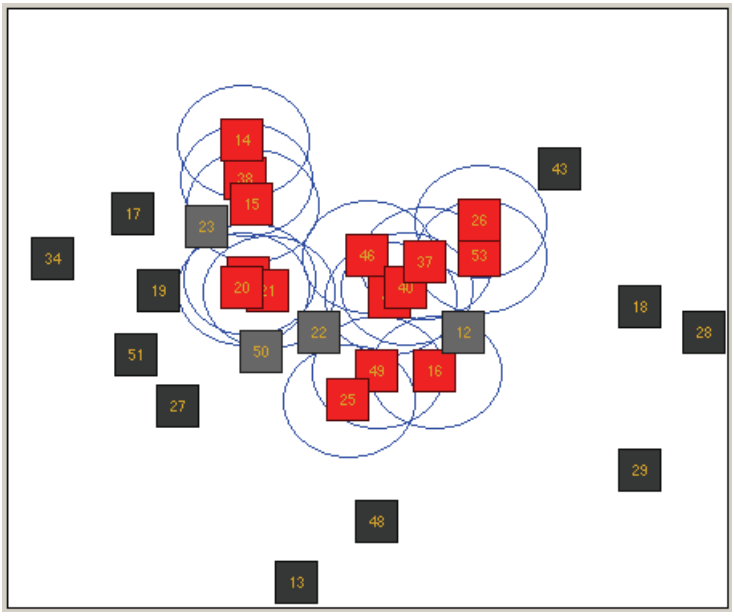


Abbildung 145: Blau eingezeichnet ist der für das jeweilige Zielitem sichtbare Horizont. Hier ist er 1/10 der maximalen Itemdistanz (mD). Rot eingefärbt sind die gefundenen Zielitems. Wieder ist die Überrepräsentation der Kartenmitte deutlich.

Noch immer ist aber die Konzentration auf die Cluster nicht zufriedenstellend. So wird keine gute Abdeckung der Karte erreicht, sondern die Clusterzentren werden überbetont.

A3.3.4 Den Einfluss des Zielitems minimieren

Eine weitere Verfeinerung des Algorithmus besteht darin, den Einfluss der Zielitems auf die benachbarten Items zu eliminieren. Die Idee dahinter ist diese, dass ein gefundener Text ja eben gefunden wurde und somit keinen Einfluss mehr auf die Nachbarn ausüben soll, sondern diese ohne dessen Hilfe genügend Kraft aufbringen müssen, um als neue Clustermitte attraktiv zu wirken. Allerdings ist damit nur eine leichte Verbesserung zu erzielen (Radius = 1/7 mD)[37].

37 Anmerkung: Wenn Items ohne Nachbarn sind, erhalten sie eine Distanzsumme von 0. Sind nur noch solche einsamen Items übrig, muss entschieden werden, welches als nächstes Zielitem genommen werden soll. Die hier gewählte Lösung bestimmte die Reihenfolge anhand der Entfernung zur Kartenmitte: Zentralere Items werden zuerst gewählt, somit werden die restlichen Items von innen nach aussen ausgewählt.

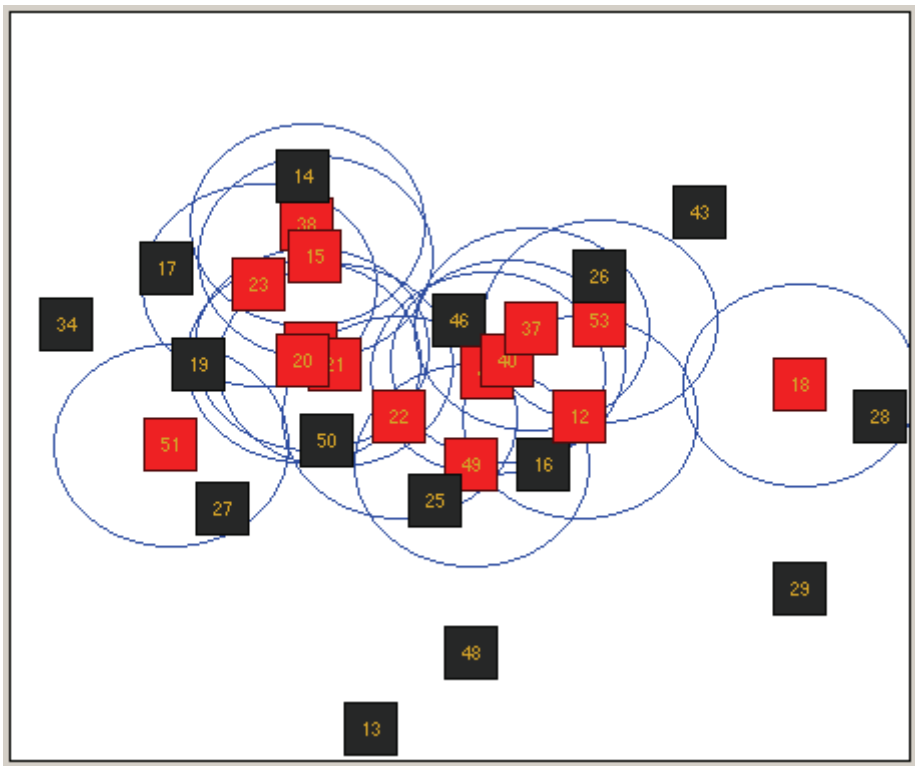


Abbildung 146: Leicht bessere Verteilung mit dem Algorithmus «Einfluss des Zielitems wird eliminiert»: Wenn ein Zielitem bestimmt wurde, hat es keinen Einfluss mehr auf die Nachbaritem. Rot eingefärbt sind die ausgewählten Zielitems, blau ist der Sichtbarkeitshorizont, mD ist 1/7.

A3.3.5 Zielitem eliminiert Nachbarn

Verfolgt man obige Idee weiter, kommen wir zum definitiven Algorithmus: Wird ein Zielitem gefunden, wird der Einfluss sämtlicher benachbarter Items eliminiert; sie kommen nicht mehr als Zielitems in Frage. Das stimmt auch mit der Anforderung überein, dass ein Zielitem einen Bereich abdecken soll.

Dieser Algorithmus funktioniert nun sehr gut. Die Sequenz zeigt in Abbildung 147 zeigt, wie die augenscheinlich relevanten Cluster erkannt werden (mD 1/9). Schon ab dem 6. Item werden sämtliche Cluster – bei gewähltem Sichtbarkeitshorizont – erkannt.

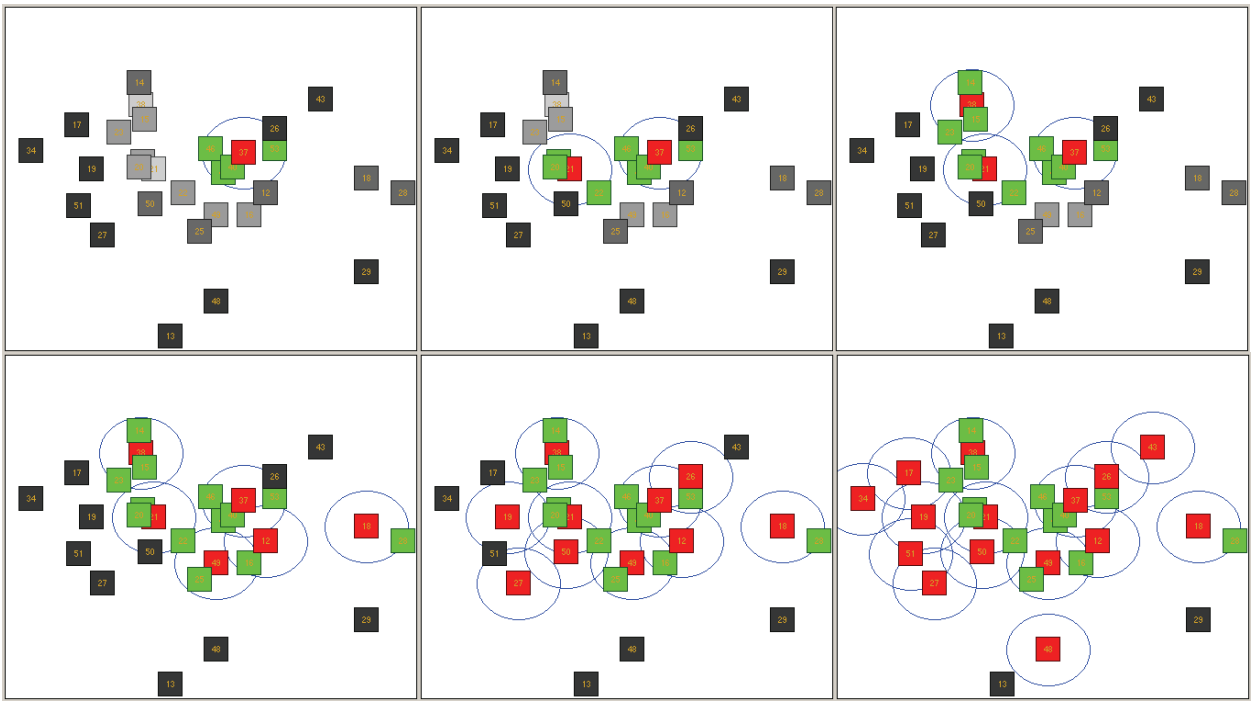


Abbildung 147: Rot: Zielitems, Grün: benachbarte Items (die nicht mehr als Zielitems in Frage kommen)

Wie sieht es inhaltlich aus? - Folgende Karte wurde aus den Ähnlichkeiten von Symposiumsbeiträgen generiert. Links ist die Einfärbung der 10 verschiedenen Symposiumsthemen eingezeichnet, rechts die gefundenen Zielitems. Tatsächlich werden die eng geclusterten Themen alle gefunden.

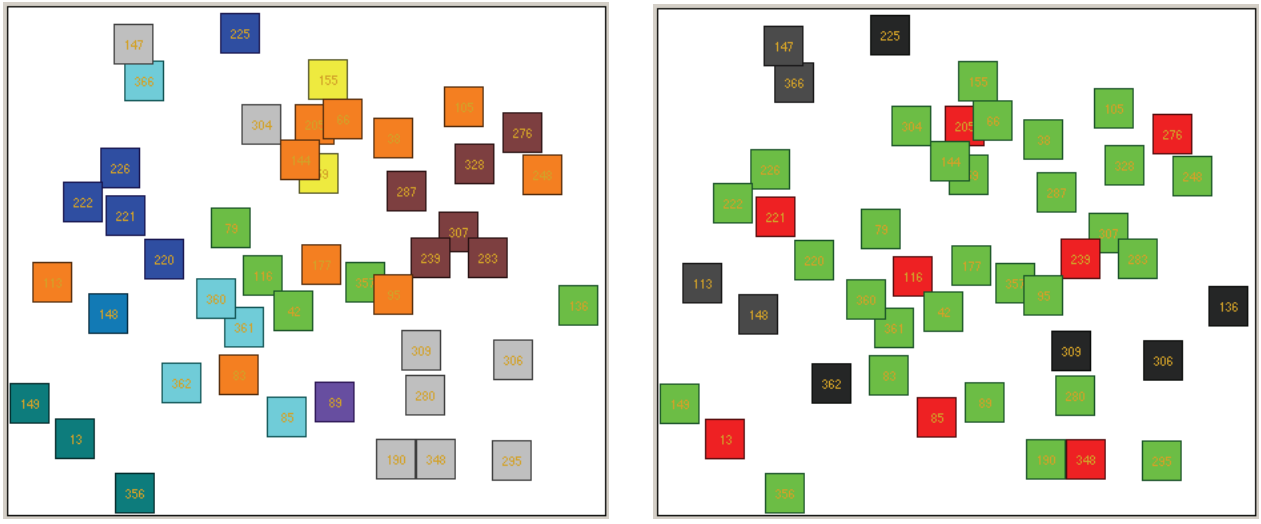


Abbildung 148: Links sind die 10 Symposiumsthemen eingezeichnet, rechts die gefundenen Zielitems (8 Zielitems, rot eingefärbt, mD 1/7; grün eingefärbt sind die Items innerhalb des Sichtbarkeitshorizontes eines Zielitems). Die eng geclusterten Themen wurden alle identifiziert.

A3.4 Diskussion

In diesem Kapitel wurde schrittweise ein Algorithmus entwickelt, der basierend auf der Verteilung von Items in einer Karte die Zielitems auswählt. Der Algorithmus nimmt zwei Parameter entgegen: Anzahl Zielitems und Nachbarradius/Sichtbarkeitshorizont. Die Bestimmung beider Parameter wurde nicht ausgetestet. Die Parameter können durch folgende Verfahren bestimmt sein.

Anzahl Zielitems:

(i) fixer Wert
(ii) abhängig von der Gesamtanzahl Items
(iii) abhängig von der Verteilung der gefundenen Zielitems (welche Bereiche der Karte abgedeckt werden)
(iv) abhängig von der Abdeckung der gefundenen Zielitems (welche Items innerhalb des Nachbarradius' liegen)

Nachbarradius/Sichtbarkeitshorizont:

(i) fixer Wert
(ii) abhängig von der Maximaldistanz zweier Items
(iii) abhängig von der Durchschnittsdistanz zweier Items

Welches Verfahren am geeignetsten ist, ist vom Zweck abhängig und wird im Rahmen dieses Kapitels nicht untersucht.

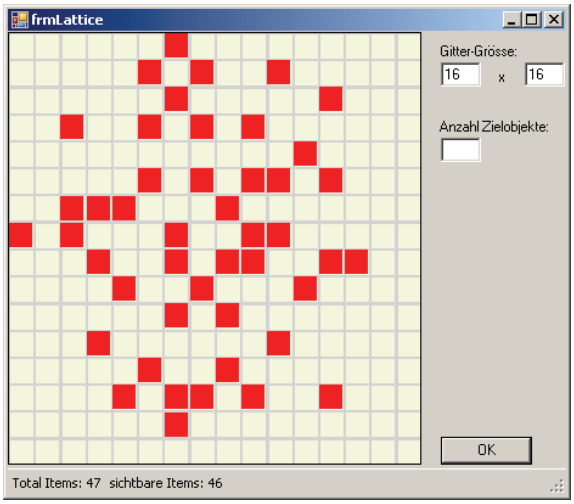


Abbildung 149: Zur Berechnung der Prototypen war erst die Idee eines Zellulären Automaten. Jedoch erwies sich der programmierte Prototyp als unpraktikabel, weil die verschiedenen Bedeutungsbereiche zur Kettenbildung neigten.

Literaturverzeichnis

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2000). *Multivariate Analysemethoden*. Berlin: Springer.

Baumeister, R. F. (1991). *Meanings of life*. New York, NY: Guilford Press.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling. Theory and Applications*. New Yord: Springer.

Borg, I., Groenen, P., & Mair, P. (2010). *Multidimensionale Skalierung*. München: Rainer Hampp Verlag.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5., vollst. überarb. und aktualisierte Aufl. ed.). Berlin: Springer-Verlag.

Danowski, J. A. (2011). Counterterrorism Mining for Individuals Semantically-Similar to Watchlist Members. In U. K. Wiil (Ed.), *Counterterrorism and Open Source Intelligence*. Wien: Springer-Verlag.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Dhillon, I. S., Modha, D. S., & Spangler, W. S. (1998). *Visualizing Class Structure of Multidimensional Data.* Paper presented at the Proceedings of the 30th Symposium on the Interface: Computing Science and Statistics, Minneapolis.

Dilling, H. (1994). *Internationale Klassifikation psychischer Störungen*. Bern: Huber.

Dilling, H. (2000). *Die vielen Gesichter des psychischen Leids: Das offizielle Fallbuch der WHO zur ICD-10 Kapitel V(F): Falldarstellungen von Erwachsenen*. Bern: Huber.

Egli, S., Schlatter, K., Streule, R., & Läge, D. (2006). A Structure-Based Expert Model of the ICD-10 Mental Disorders. *Psychopathology, 39*, 1–9.

Freyberger, H. J., & Dilling, H. (1999). *Fallbuch Psychiatrie: Kasuistiken zum Kapitel V (F) der ICD-10*. Bern: Huber.

Hörmann, H. (1976). *Meinen und Verstehen*. Frankfurt am Main: Suhrkamp.

Joachims, T. (2001). *A Statistical Learning Model of Text Classification for Support Vector Machines*. Paper presented at the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Krieg, P. (2005). *Die paranoide Maschine - Computer zwischen Wahn und Sinn*: Heise.

Lieb, K., Hesslinger, B., & Jacob, G. (2009). *50 Fälle Psychiatrie und Psychotherapie: Bed-side-learning*. München: Urban & Fischer Verlag/Elsevier GmbH.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28*(2), 203-208.

Marx, W. (1976). Die Messung der Assoziativen Bedeutungsähnlichkeit. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 23(1), 62–76.

Marx, W., & Hejj, A. (1989). *Subjektive Strukturen Ergebnisse aus der Gedächtnis-, Sprach- und Einstellungsforschung*. Göttingen: Hogrefe.

McCallum, A., & Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. Technical Report. Workshop on Learning for Text Categorization, 41-48.

Mathar, R. (1997). *Multidimensionale Skalierung*. Stuttgart: Teubner.

Maturana, H. R. & Varela, F. J. (1990). *Der Baum der Erkenntnis:* Goldmann.

Michel, O. (2006). *Die Hofmethode: Auf dem Weg zum maschinellen Textverständnis.* Paper presented at the Usability Day IV, Vorarlberg.

Michel, O., & Läge, D. (2007). *The Congress Map: Documenting the similarity of conference contributions by information retrieval from the scientific abstracts.* Paper presented at the Kongress SGP 2007, Zürich.

Michel, O., & Läge, D. (2009). The Hofmethode: Computing semantic similarities between e-learning products, *Interactive Computer Aided Learning (ICL) 2009*. Villach, Austria.

Michel, O., & Läge, D. (2009). The Hofmethode: Computing Semantic Similarities between E-Learning Products, *International Journal of Emerging Technologies in Learning (iJET)* (Vol. 4).

Michel, O., & Läge, D. (2010). *Semantic Structuring of Conference Contributions Using the Hofmethode.* Paper presented at the I-KNOW 2010, Graz, Austria.

Miller, G. (1956). The Magical Number 7, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63.

Osgood, C. E. (1952). The Nature and Measurement of Meaning. *Psychological Bulletin, 49*, 197-237.

Osgood, C. E. (1963). On Understanding and Creating Sentences. *American Psychologist, 18*, 735-751.

Osgood, C. E., & Tzeng, O. C. S. (1990). *Language, meaning, and culture: The selected papers of C. E. Osgood.* New York: Praeger.

Panyr, J. (1986). *Automatische Klassifikation und Information Retrieval Anwendung und Entwicklung komplexer Verfahren in Information-Retrieval-Systemen und ihre Evaluierung*. Tübingen: Niemeyer.

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Massachusetts: Massachusetts Institute of Technology.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*, 81–106.

Ryf, S., & Läge, D. (2008). Berechnung und Visualisierung von Verteilungen in NMDS-Karten am Beispiel des Musik- und Getränkemarktes. In J. Reinecke & C. Tarnai (Eds.), K*lassifikationsanalysen in Theorie und Praxis*. Münster: Waxmann Verlag.

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613–620.

Sass, H., & Houben, I. (1998). *Diagnostisches und statistisches Manual psychischer Störungen DSM-IV : übersetzt nach der vierten Auflage des Diagnostic and statistical manual of mental disorders der American Psychiatric Association*. Göttingen: Hogrefe.

Schneider, M. (2009). *Theorie der Semantik*, from http://www.semager.de/info/theoretische_semantik.ppt

Spangler, W. S., Kreulen, J. T., & Newswanger, J. F. (2006). Machines in the conversation: Detecting themes and trends in informal communication streams. *IBM Systems Journal, 45*(4), 785–799.

Wiil, U. K. (Ed.). (2011). *Counterterrorism and Open Source Intelligence*. Wien: Springer.

Wittgenstein, L. (1960). *Philosophische Untersuchungen*. Frankfurt.

Die amerikanischen Journalisten beschrieben die tanzenden Sklavinnen Hollywoods. Einer fing an, den Elvis Presley-Song zu pfeifen, den Elvis singt, während er in *Harum Scarum* (1965) in einen Harem eindringt:

„I am gonna go where the desert sun is, where the fun is,
go where the harem girls dance,
go where there's love and romance ...

Ich gehe in die Wüste, dort wartet das Glück,
wo die Haremsfrauen tanzen,
wo sie mich lieben und küssen ...[10]

Gekleidet wie ein Araber reitet Elvis Presley pfeilschnell durch die Wüste, stürzt sich in einen Harem und befreit eine dort verbor-

aus Fatima Mernissi: Westliche Phantasien, Harem, östliche Wirklichkeit